Novel approaches to analysing 16S ribosomal RNA gene sequencing profiles of the gut microbiota in human health research.

Jackson, Matthew Albert

Awarding institution:
King's College London

Novel approaches to analysing 16S ribosomal RNA gene sequencing profiles of the gut microbiota in human health research.

Matthew A. Jackson

A thesis presented for the degree of
Doctor of Philosophy

Department of Twin Research and Genetic Epidemiology
King's College London
United Kingdom
2017

Abstract

There is an increasing body of research associating the gut microbiome with human health. This has been fuelled by the development of untargeted techniques to profile whole microbial communities. A common approach is to sequence the 16S rRNA gene as a marker for taxonomic quantification. However, the results of such analyses are of limited use. The high-dimensional sequencing data are typically reduced to operational taxonomic units (OTUs), which are purely analytical units with no direct biological interpretation. Furthermore, experimental and analytical variation can lead to poor reproducibility of results between such studies. Here, I present novel methods to address these limitations and improve the application of 16S rRNA gene sequencing to profiling of the gut microbiota in human health research.

I first explore existing and simpler approaches to the analysis of 16S rRNA gene sequencing data by investigating gut microbiota associations with two phenotypes, host frailty and proton pump inhibitor use, identifying several novel associations with both. I then tackle the limited biological representation of OTUs by presenting a comprehensive comparison of OTU clustering algorithms, using heritability as a novel and biologically motivated quality measure. The results of this comparison provide guidance for the approaches used in later analyses. Finally, I present two novel, more complex, methods to summarise 16S rRNA gene sequencing data in human health studies; both of which address the high dimensionality and limited reproducibility and biological relevance of these datasets. Firstly, I identify gut microbiota associations with multiple diseases within a single study. This generates comparable results that enable the identification of key marker taxa consistently associated with health or disease. From this, I generate an index that can represent wide-scale gut microbiota composition within a single score that is also associated with host health. Secondly, I present a method to identify communities of interacting microbes within the gut microbiota and demonstrate that these reliably associate with host phenotypes across geographically diverse populations.

In summary, this thesis addresses several issues currently associated with the analysis of 16S rRNA gene sequencing-based gut microbiota profiles and in turn identifies several novel biological phenomena. The techniques described herein should improve the utility of microbiota profiles derived from 16S rRNA gene sequencing and advance the field of gut microbiome research in human health.

Acknowledgements

I would first like to thank my three supervisors, Claire Steves, Jordana Bell, and Tim Spector, for their help throughout my PhD. In particular, I would like to thank Jordana for help with developing my writing skills and career advice; Tim for providing access to a diverse and wide range of data and collaborators that enabled me to take part in numerous large-scale projects as part of, and alongside, my PhD; and Claire for allowing me the freedom to follow my own ideas, giving support and steering when it was needed most, and providing my first insights into the clinical world. I would also like to thank Francis Williams, Kerrin Small, and Kevin Whelan. Whilst, not directly supervising my projects, all three provided guidance in some way that ultimately improved the direction of my work.

I would also like to thank the collaborators that have directly contributed to work within this thesis. From TwinsUK these include Jonas Zierer, Ruth Bowyer, Serena Verdi, and Cheol Min Shin amongst others. I have also been fortunate to work with many collaborators from a number of different institutions. In particular I would like to thank Julia Goodrich, Ruth Ley, Andrew Clark, Paul O'Toole, Ian Jeffery, Daniel Freedberg, Maria-Emanuela Maxan, Zhana Kuncheva, and Marc Jan Bonder. All of whom have provided invaluable work and/or advice that has contributed in some way to the projects in this thesis.

I would also like to thank all the other members of the department who work behind the scenes to keep things running and have proved to become friends over the last three years. Whilst a PhD can inevitably be fraught at times, it is made much easier working in such a welcoming environment. Particular thanks go to Antonino Zito and Juan Castillo-Fernandez, I hope we have many more beers (only halves for Antonino) in the years to come. I must also thank the twins themselves. These volunteers provide a valuable service and this thesis would not exist without them.

Thank you also to my friends and family that have provided support not only in the last three years but in all the time that has led me here. Thanks especially to my parents, I'm sure it hasn't been cheap, I hope you enjoy reading this thesis. Finally, I would like to thank my flatmate, best friend, and partner Sophie Wood. She, more than anyone, has maintained my sanity throughout all the pressures of the last three years; whilst I have tested hers with proof-reading, practice talks, and poorly-timed discussions about faeces. Thanks for not only putting up with me but supporting me, it really was necessary.

Publications

- **Jackson MA**, Goodrich JK, Maxan ME, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD. Proton pump inhibitors alter the composition of the gut microbiota. Gut. 2015 Dec 30.
- **Jackson MA**, Jeffery IB, Beaumont M, Bell JT, Clark AG, Ley RE, O'Toole PW, Spector TD, Steves CJ. Signatures of early frailty in the gut microbiota. Genome Medicine. 2016 Jan 29.
- Goodrich JK, Davenport ER, Beaumont M, **Jackson MA**, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. Genetic determinants of the gut microbiome in UK twins. Cell Host & Microbe. 2016 May 11.
- **Jackson MA**, Bell JT, Spector TD, Steves CJ. A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. PeerJ. 2016 Aug 30.
- Beaumont M, Goodrich JK, **Jackson MA**, Yet I, Davenport ER, Vieira-Silva S, Debelius J, Pallister T, Mangino M, Raes J, Knight R. Heritable components of the human fecal microbiome are associated with visceral fat. Genome Biology. 2016 Sep 26.
- Xie H, Guo R, Zhong H, Feng Q, Lan Z, Qin B, Ward KJ, **Jackson MA**, Xia Y, Chen X, Chen B. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. Cell Systems. 2016 Dec 21.
- Menni C, **Jackson MA**, Pallister T, Steves CJ, Spector TD, Valdes AM. Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain. International Journal of Obesity. 2017 Jul.
- Pallister T, **Jackson MA**, Martin TC, Glastonbury CA, Jennings A, Beaumont M, Mohney RP, Small KS, MacGregor A, Steves CJ, Cassidy A. Untangling the relationship between diet and visceral fat mass through blood metabolomics and gut microbiome profiling. International Journal of Obesity. 2017 Jul.
- Le Roy CI, Beaumont M, **Jackson MA**, Steves CJ, Spector TD, Bell JT. Heritable components of the human fecal microbiome are associated with visceral fat. Gut Microbes. 2017 Jul 19.
- Menni C, Zierer J, Pallister T,**Jackson MA**, Long T, Mohney RP, Steves CJ, Spector TD, Valdes AM. Omega-3 fatty acids correlate with gut microbiome diversity and production of N-carbamylglutamate in middle aged and elderly women. Scientific Reports. 2017 Sep 11.
- Pallister T, **Jackson MA**, Martin TC, Zierer J, Jennings A, Mohney RP, MacGregor A, Steves CJ, Cassidy A, Spector TD, Menni C. Hippurate as a metabolomic marker of gut microbiome diversity: Modulation by diet and relationship to metabolic syndrome. Scientific Reports. 2017 Oct 20.
- **Jackson MA**, Bonder MJ, Kuncheva Z, Zierer J, Fu J, Kurilshikov A, Wijmenga C, Zhernakova A, Bell JT, Spector TD, Steves CJ. Detection of stable community structures within gut microbiota co-occurrence networks from different human populations. PeerJ. 2018 Feb 7.

Contents

Abstract **2**

Acknowledgements **3**

Publications **4**

1 Introduction **14**
1.1 Overview . . . 14
1.2 The human gut microbiome . . . 15
1.2.1 The human microbiome . . . 15
1.2.2 The gut microbiome in disease . . . 15
1.2.3 Mechanisms underlying gut microbiota-disease associations . . . 21
1.2.4 Host determinants of gut microbiome composition . . . 27
1.3 16S rRNA gene profiling of microbiota . . . 31
1.3.1 Phylogenetics and the ribosome . . . 31
1.3.2 Sequencing 16S rRNA genes in the gut microbiome . . . 32
1.3.3 Operational Taxonomic Units . . . 35
1.3.4 Typical approaches to the analysis of 16S rRNA gene sequencing-based gut microbiome profiles . . . 41
1.3.5 Alternative approaches to assay the gut microbiome . . . 45
1.4 Co-occurrence networks in the gut microbiome . . . 46
1.4.1 Microbial interactions . . . 46
1.4.2 Quantification of co-occurrence networks from 16S rRNA gene sequencing profiles of the gut microbiota . . . 48
1.4.3 Defining microbial communities from co-occurrence networks . 50
1.5 Aims . . . 54

2 Materials and Methods **55**
2.1 TwinsUK . . . 55
2.1.1 TwinsUK database . . . 55
2.1.2 Heritability estimation . . . 57
2.2 16S rRNA gene sequencing from faecal samples . . . 57
2.2.1 Sample collection and processing . . . 57

2.2.2 Data batches and processing . . . 58
2.3 Other microbiome datasets . . . 59
2.3.1 Metagenomics . . . 59
2.3.2 Faecal metabolomics . . . 60

3 Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with host frailty in the gut microbiota 61

4 Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with proton pump inhibitor use in the gut microbiota 73

5 A comparison of methods to cluster 16S rRNA gene sequences to OTUs using heritability as a measure of quality 82

6 Identification of taxonomic markers that define a health-associated gut microbiome 102
6.1 Introduction . . . 103
6.2 Methods . . . 104
6.2.1 Disorder and drug data . . . 104
6.2.2 Processing of 16S rRNA gene sequences . . . 104
6.2.3 Greedy selection of marker traits . . . 107
6.2.4 Regression analyses . . . 108
6.2.5 Inference of a Bayesian network between disorders, drug use, and microbiome marker traits . . . 109
6.2.6 Clustering of microbiome traits by disorder associations . . . 109
6.2.7 Calculating the HMI and MDI . . . 110
6.2.8 Index associations with TwinsUK microbiota and health . . . 110
6.2.9 Optimising the HMI . . . 110
6.2.10 Replication datasets . . . 111
6.2.11 Index associations in replication data . . . 112
6.2.12 Host influences on the HMI . . . 112
6.2.13 Metagenomic and metabolomic enrichment analysis . . . 113
6.3 Results . . . 114
6.3.1 Common disorders and medications in TwinsUK . . . 114
6.3.2 Disorder and drug associations with marker traits in the gut microbiome . . . 116
6.3.3 Delineating disorder-drug effects . . . 120
6.3.4 Defining a health-associated gut microbiome . . . 122
6.3.5 The health-associated microbiome index (HMI) . . . 124
6.3.6 The robustness of the HMI . . . 126
6.3.7 Host influences on the health-associated gut microbiome . . . 131
6.3.8 Functionality of the health-associated gut microbiome . . . 134

6.4 Discussion . . . 135
6.4.1 Associations with gut microbiota markers . . . 135
6.4.2 A health associated gut microbiota . . . 136
6.4.3 The healthy microbiome index . . . 137
6.4.4 Functions of a health-associated gut microbiota . . . 137
6.4.5 Conclusion . . . 138

7 Identifying stable communities in the gut microbiota with consistent associations to host phenotypes 140
7.1 Introduction . . . 141
7.2 Methods . . . 142
7.2.1 Data aggregation . . . 142
7.2.2 OTU clustering, table filtering, and diversity analyses . . . 144
7.2.3 An ensemble approach to co-occurrence networks . . . 145
7.2.4 Data-driven determination of a p-value threshold . . . 146
7.2.5 Detecting communities within co-occurrence networks . . . 146
7.2.6 Community properties and host associations . . . 148
7.2.7 Comparison of community structure between datasets . . . 148
7.3 Results . . . 149
7.3.1 Overlap of the TwinsUK, LLDEEP, and Israeli-PN data . . . 149
7.3.2 Scale-free thresholding of edges in ensemble networks . . . 150
7.3.3 Detection of communities in co-occurrence networks . . . 154
7.3.4 Taxonomy, heritability, and host associations of communities in TwinsUK . . . 157
7.3.5 Community structure and associations across populations . . . 159
7.4 Discussion . . . 162
7.4.1 Differences in the gut microbiota between datasets . . . 162
7.4.2 Generating interaction networks . . . 163
7.4.3 Robust community detection . . . 163
7.4.4 Stable community structure across populations . . . 164
7.4.5 Conclusion . . . 165

8 General Discussion 166
8.1 Summary of results . . . 166
8.1.1 Methodological Findings . . . 166
8.1.2 Biological Findings . . . 169
8.2 Limitations and Future Work . . . 171
8.2.1 Use of observational cohort data . . . 171
8.2.2 Wider application of results . . . 174
8.2.3 Alternate approaches to microbiome profiling . . . 176
8.3 Conclusion . . . 178

References **179**

Appendices **195**

A Chapter 3: Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with host frailty in the gut microbiota **196**

B Chapter 4: Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with proton pump inhibitor use in the gut microbiota **200**

C Chapter 5: A comparison of methods to cluster 16S rRNA gene sequences to OTUs using heritability as a measure of quality **222**

D Chapter 6: Identification of taxonomic markers that define a health-associated gut microbiome **232**

E Chapter 7: Identifying stable communities in the gut microbiota with consistent associations to host phenotypes **248**

List of Figures

1.1 Demonstration of the taxonomic diversity of the gut microbiota and its variability between individuals. . . . 18
1.2 Connectivity of influences on the human gut microbiome. . . . 31
1.3 Diagrammatic representation of the secondary structure of an *Escherichia coli* 16S rRNA. . . . 33
1.4 Overview of an experimental pipeline to generate 16S-based gut microbiota profiles at a large population cohort scale. . . . 35
1.5 Schematic of the idealised concept of OTU clustering. . . . 37
1.6 Overview of a greedy approach to clustering read data based on the kmer indexing approach used by VSEARCH. . . . 38
1.7 Two taxonomy based diversity measures used in microbiome analyses. . . . 43
1.8 Summary of different microbial interaction types with examples. . . . 47
1.9 Equation for the log ratio transform. . . . 49
1.10 Flowchart for selecting the appropriate co-occurrence method to use for microbiome data. . . . 50
1.11 The equation to estimate modularity from a partitioned network. . . . 52

2.1 Summary properties of the 2764 individuals with 16S rRNA gene sequencing of their gut microbiota within TwinsUK. . . . 56
2.2 Path diagram specifying a structural equation model for a typical ACE model. 58

3.1 OTU abundances significantly different between twin pairs discordant for frailty. . . . 66
3.2 Species abundance associates with FI. . . . 67
3.3 Correlation of frailty, diversity, and model covariates with modules of OTUs collapsed by co-occurrence. . . . 69

4.1 Distributions of covariates included for analysis, compared between proton pump inhibitor (PPI) users and non-users. . . . 77
4.2 Comparison of α-diversity in proton pump inhibitor (PPI) users and non-users, and in individuals with and without GI indications. . . . 78
4.3 Summary of taxonomic associations with proton pump inhibitor (PPI) use. 79

4.4 Paired plots of the relative abundances of Streptococcaceae and Lactobacillales within monozygotic (MZ) twins discordant for proton pump inhibitor (PPI) use. 80

5.1 Twin based A, C, and E estimate comparisons between closed and open reference, and de novo clustering using UCLUST with a similarity threshold of 97%. 88
5.2 Twin based A, C, and E estimate comparisons between different greedy algorithms for de novo clustering at a 97% similarity threshold. 90
5.3 Twin based A, C, and E estimate comparisons between three different thresholds of de novo clustering using VSEARCH, VSEARCH de-replicated sequences, and two non-threshold based techniques. 91
5.4 Comparison of A heritability estimates between all clustering approaches. Only considering OTUs who's A estimate was greater than the mean (8%) and had a lower 95% CI greater than 1%. 92
5.5 Comparison of absolute alpha diversity values for Shannon, Simpson, Chao1, and OTU count indices across all samples. 93
5.6 Kendall's Tau rank based correlations between samples across methods for each of Shannon, Simpson, Chao1, and OTU count metrics. 94

6.1 Scree plot showing the percentage variation explained by each axis in PCoA of the weighted and unweighted UniFrac distances for all twins in the analysis. 107
6.2 Graphical example of the approach used to select marker traits in the gut microbiota. 108
6.3 Summary of the common disorders considered in the analysis. 115
6.4 Summary of the common medications used in the cohort and considered in the analysis. 116
6.5 Associations between diseases and drugs and the microbiota markers. . . 118
6.6 Number of cases compared to the number of associations observed. . . . 119
6.7 Bayesian network inferred from disorder, drug-use, and gut microbiome data. 121
6.8 Defining a universal health-associated gut microbiome. 123
6.9 Beta diversity plots showing the first two axes from PCoA analyses using the weighted UniFrac metric across all samples in TwinsUK. 125
6.10 Plots showing the association between the HMI, MDI, and Shannon indices, with the number of disorders and frailty of an individual and the Shannon diversity of their gut microbiota. 128
6.11 TwinsUK members with a *Prevotella* rich microbiome have low diversity but are significantly healthier. 129
6.12 Health and wide-scale microbiota composition associations of the HMI replicate in external datasets. 130

6.13 Using the HMI to identify host influences on, and functional properties of, the health-associated microbiome. 133

7.1 An outline of the study design used in this chapter. 143
7.2 Diveristy comparisons between datasets. 150
7.3 Taxonomic similarity of the datasets considered. 151
7.4 An example visualisation of the ensemble process at the p<0.01 threshold for TwinsUK. 152
7.5 Selection of final p-value threshold for generating ensemble networks. . . 153
7.6 Selecting a γ parameter for Louvain thresholding. 155
7.7 A visualisation of the communities detected by the Louvain algorithm. . . 156
7.8 Summary of the taxonomic distributions within the TwinsUK communities. 157
7.9 Host associations with TwinsUK communities. 158
7.10 Pair-wise comparison of the variation of information between community definitions for OTUs shared between networks. 159
7.11 Mapping communities across datasets to identify stable community-types. 161
7.12 Replication of linear regression analysis for each of the 14 mapped community-types in the LLDEEP and Israeli-PN datasets. 162

D.1 Questionnaire time differences from the faecal sample dates. 232
D.2 Disorder associations with microbiota markers without adjustment for BMI. 240

List of Tables

1.1 Definitions for some common terms related to microbiome research that are used within this thesis. . . . 16
1.2 A comparison of OTU clustering approaches. . . . 40

2.1 Batches of 16S rRNA gene sequencing data produced within TwinsUK. . 59

3.1 Alpha diversity associations with covariates and the frailty index. . . . 65

6.1 The 38 common disorders (>1% of TwinsUK) considered. . . . 105
6.2 The 51 common medications (>1% of TwinsUK) considered. . . . 106
6.3 PERMANOVA results for index associations with UniFrac measures in TwinsUK. . . . 124
6.4 The familial marker traits used to define the optimised HMI. . . . 127
6.5 PERMANOVA results for index associations with UniFrac measures in replication datasets. . . . 131
6.6 Results of association analyses between dietary nutrient intakes and the HMI using two different nutrient summary levels. . . . 134

7.1 Participant summary statistics for the datasets considered. . . . 149
7.2 Summary statistics for the final intersect co-occurrence networks ($p<0.01$) for each dataset. . . . 154

A.1 **S1.** Table of domains used in the construction of the frailty index. . . . 196
A.2 **S3.** Table of of significant OTU-FI associations. . . . 197
A.3 **S4.** Table of significant taxonomic associations with the FI. . . . 198
A.4 **S5.** Table of OTU associations with FI in non-parametric analyses. . . . 199
A.5 **S6.** Table of OTU module assignments. . . . 199

B.1 **S1.** Summary of the aspects covered by the health domains considered in the construction of the frailty index. . . . 203
B.2 **S2.** Table of alpha diversity associations with PPI use after adjustment for covariates. . . . 203
B.3 **S3.** Significant results from modelling OTU associations with PPI use. . . . 204

B.4 **S4.** Significant taxa associations with PPI use and replication results in the interventional data set. . . . 206
B.5 **S5.** Significant results from mixed effects modelling of PPI and OTU and taxa associations in the subset of individuals with complete covariate data and no antibiotic use within the previous month. . . . 211
B.6 **S6.** Table of HMP derived site specificities for collapsed families. . . . 216
B.7 **S7.** Table showing HMP site definition mappings to the broader categories used in this study. . . . 220

C.1 **S1.** Summary of OTU counts resulting from each method . . . 231
C.2 **S2.** Method-wise A,C and E estimates for each taxa found across all methods. . . . 231

D.1 Mapping of the 206 gut microbiome features considered to their marker in the 68 marker features used in analyses. . . . 233
D.2 Results from logistic regression analysis of disorders versus microbiome features. . . . 240
D.3 Results from logistic regression analysis of disorders versus microbiome features without adjustment for BMI. . . . 240
D.4 Edge table for the inferred Bayesian network underlying the disorder, drug and microbiome data. . . . 240
D.5 Enrichment analysis results for KEGG annotated pathways from metagenomic data against the HMI, number of disorders, and frailty. . . . 240
D.6 Enrichment analysis results for KEGG annotated pathways from faecal metabolite data against the HMI, number of disorders, and frailty. . . . 241

E.1 OTU community assignments and heritability analysis results for the communities in TwinsUK. . . . 248
E.2 Community association results with age and BMI in TwinsUK (Community members can be found in Table E.1) . . . 248
E.3 Table summarising the OTUs within each of the community types identified across all three datasets. . . . 250

Chapter 1

Introduction

1.1 Overview

Within this thesis I present novel approaches to the analysis of 16S ribosomal RNA (rRNA) gene sequencing-based profiles of the human gut microbiota. With the aim to improve their applicability to human health research. This is achieved through:

1. Benchmarking one of the key stages in 16S rRNA gene sequencing analysis - the creation of operational taxonomic units.
2. The identification of marker taxa, and the development of an associated index, to quantify gut microbiome composition in relation to host health.
3. The development of a robust approach to identify microbial communities in the gut microbiome.

These sections are also preluded by analyses of two health related phenotypes, host frailty and proton pump inhibitor (PPI) use, as a demonstration of typical approaches to the analysis of 16S rRNA gene sequencing in microbiome research.

In the present chapter I will introduce the concept of the human microbiome and provide a brief overview of the gut microbiome and its associations with human health. I will then describe the process of microbiota profiling by 16S rRNA gene sequencing, preparing data for analysis, the concept of operational taxonomic units, and typical methods used to analyse microbiome data in relation to human health; highlighting some of the limitations of these approaches throughout. I will then examine the existing methods used to identify microbial communities in 16S rRNA gene sequencing profiles of the gut microbiome.

Following this chapter is a description of the data and methods used throughout. As this thesis is presented through publication, the first three of the five research chapters are presented as manuscripts unmodified from their published form. I conclude with a summary of the thesis in its entirety.

1.2 The human gut microbiome

1.2.1 The human microbiome

The term microbiome refers to the ecological community of microbes (chiefly bacteria, archaea, and fungi) inhabiting a particular environment (Table 1.1). This terminology is now applied to studies of microbial communities from a wide range of environments but was originally coined in reference to those inhabiting the human body (Lederberg and McCray 2001). Humans are host to trillions of microbes found in communities at a range of sites across the body, including gut, oral, nasal, skin, and vaginal microbiomes (Consortium 2012). Across the body the number of bacterial cells is estimated to be approximately equal to the number of human cells but contains a much vaster genetic repertoire (Sender, Fuchs, and Milo 2016). Whilst there has been previous research of the human microbiome using culture based approaches, it is only more recently through the development of culture independent techniques (see Section 1.3) that it has been possible to examine the full diversity and functionality of these systems. This has lead to a rapid expansion in microbiome research (Jones 2013).

Costello *et al.* carried out one of the first studies using culture independent techniques to investigate the human microbiome across different body sites and between individuals. Investigating six body sites in nine individuals, they demonstrated that microbiome composition primarily segregated by body site with each site having a distinct microbial composition (Costello et al. 2009). Furthermore, they showed that whilst there is some temporal variability, microbiomes also retain compositional properties unique to individuals. This was later confirmed in a study investigating finer-scale temporal variation using daily sampling from three body sites (Caporaso et al. 2011b). These findings were also reproduced in one of the landmark studies of the human microbiome - the Human Microbiome Project (HMP). The project profiled the microbiomes of over 200 healthy individuals across five body areas with longitudinal sampling. The HMP not only confirmed the aforementioned findings, but has also provided a key reference for typical microbiomes at a range of body sites. It was also one of the first large-scale studies to probe microbiome function, showing there is higher inter-individual conservation of microbiome function than taxonomic composition (Consortium 2012).

1.2.2 The gut microbiome in disease

Often referred to as the gut microbiome, the colon contains the most densely populated microbial community in the body (Savage 1977). As such, it is the largest interface between the host and its microbial inhabitants. The gut microbiome can also be sampled non-

Table 1.1: Definitions for some common terms related to microbiome research that are used within this thesis. (Marchesi and Ravel 2015).

Term	Definition
Microbiome	Refers to the complete ecosystem of microbial organisms inhabiting a given environment, including its taxonomic make-up and functional properties.
Microbiota	Refers to the microbial taxa within a microbiome. Can be used in reference to any taxonomic level from strain to kingdom. Typically assayed using targeted marker gene based approaches such as 16S rRNA gene sequencing. Assesses the taxonomic distribution of a microbiome.
Metagenome	Refers to the microbial genomic complement across a whole microbiome. Typically assayed using shotgun sequencing of microbiome DNA. Assesses the functional potential of a microbiome.
16S	Refers to the 16S rRNA, a sub-unit of the ribosome found across all domains of life. Typically referred to in the context of 16S rRNA gene sequencing, where its gene is used a taxonomic marker to assay microbiota.
OTU	Operational Taxonomic Unit. Refers to any analytical unit created by grouping taxonomically/genetically related organisms together for the purpose of simplifying analyses. Typically used in reference to grouping of closely related 16S rRNA gene sequencing reads to reduce data dimensionality in microbiota profiles.
Alpha Diversity	The ecological diversity within a microbiome sample. Gives an indication of the number of different taxa observed (richness) and how comparable their abundances are (evenness). Can be quantified using established ecological indices.
Beta Diversity	The difference in diversity between two microbiome samples. This can be in terms of the presence/absence of similar taxa and/or the similarity in the evenness of their abundances. Can be quantified using a range of distance/similarity measures, which are often used in principle co-ordinates analyses to visualise clustering of multiple microbiome samples in a study.

invasively using faecal sampling. As a result of these factors, the gut microbiome is one of the most well studied in the human body.

Several studies have characterised typical gut microbiome composition and function in large population based cohorts, these include the HMP (Consortium 2012), the Flemish Gut Project (Falony et al. 2016), the Dutch LifeLines-DEEP cohort (Zhernakova et al. 2016), the European MetaHit consortium (Qin et al. 2010), the American Gut Project (*The American Gut Project*), and the British TwinsUK cohort (Goodrich et al. 2014b). The gut microbiome is one of the most ecologically diverse (high alpha diversity) when compared to those of other body sites (Consortium 2012). Within a population there are few core taxa (bacteria found across all participants) and each individual gut microbiome has unique characteristics. However, the gut microbiome has a stable high-level taxonomic composition generally dominated by anaerobic bacteria from the phyla Bacteroidetes and Firmicutes (Qin et al. 2010) (Figure 1.1). It was observed across several human populations that gut microbiomes tended to fall into three groups referred to as enterotypes, with microbiota dominated by taxa from the *Bacteroides*, *Ruminococcus*, or *Prevotella* genera (Arumugam et al. 2011; Wu et al. 2011). It has since been debated if enterotypes are as distinct as first thought or represent peaks in a gradient of taxonomic compostionality (Jeffery et al. 2012). However, grouping of enterotypes has been replicated (Falony et al. 2016; Qin et al. 2012) and at least the existence of the Bacteroides and Prevotella dominant groups has been more widely reported (Wu et al. 2011; Claesson et al. 2012).

Alongside establishing the typical properties of the gut microbiome, much research within cohort-based and smaller targeted projects has focused on the role of the gut microbiome in human health and disease. The gut microbiome's role as an intermediary between environmental effects and its potentially malleable nature make it a promising target for manipulation in disease treatment. Here, I provide a brief overview of some well established gut microbiome-disease associations. The potential mechanisms underlying these disease-gut microbiota associations are discussed in Section 1.2.3.

Inflammatory bowel disease and irritable bowel syndrome

One of the most studied diseases in relation to the gut microbiome is inflammatory bowel disease (IBD). This is an encompassing term for diseases of chronic inflammation in the GI tract including Crohn's disease (CD) and Ulcerative Colitis (UC). There is a known genetic and immune component to these diseases but they have also been associated with changes in the gut microbiota (Ananthakrishnan 2015). CD is associated with a decreased microbial diversity in the gut microbiome accompanying a decrease in the abundance of the butyrate producing bacteria *Faecalibacterium prausnitzii* (Sokol et al. 2008; Manichanh et al. 2006). A study of over 400 untreated paediatric patients with CD identified several taxonomic signatures that were sufficient to diagnose patients compared to healthy controls (Gevers et al. 2014). Similarly there is a loss of ecological diversity in the gut microbiome

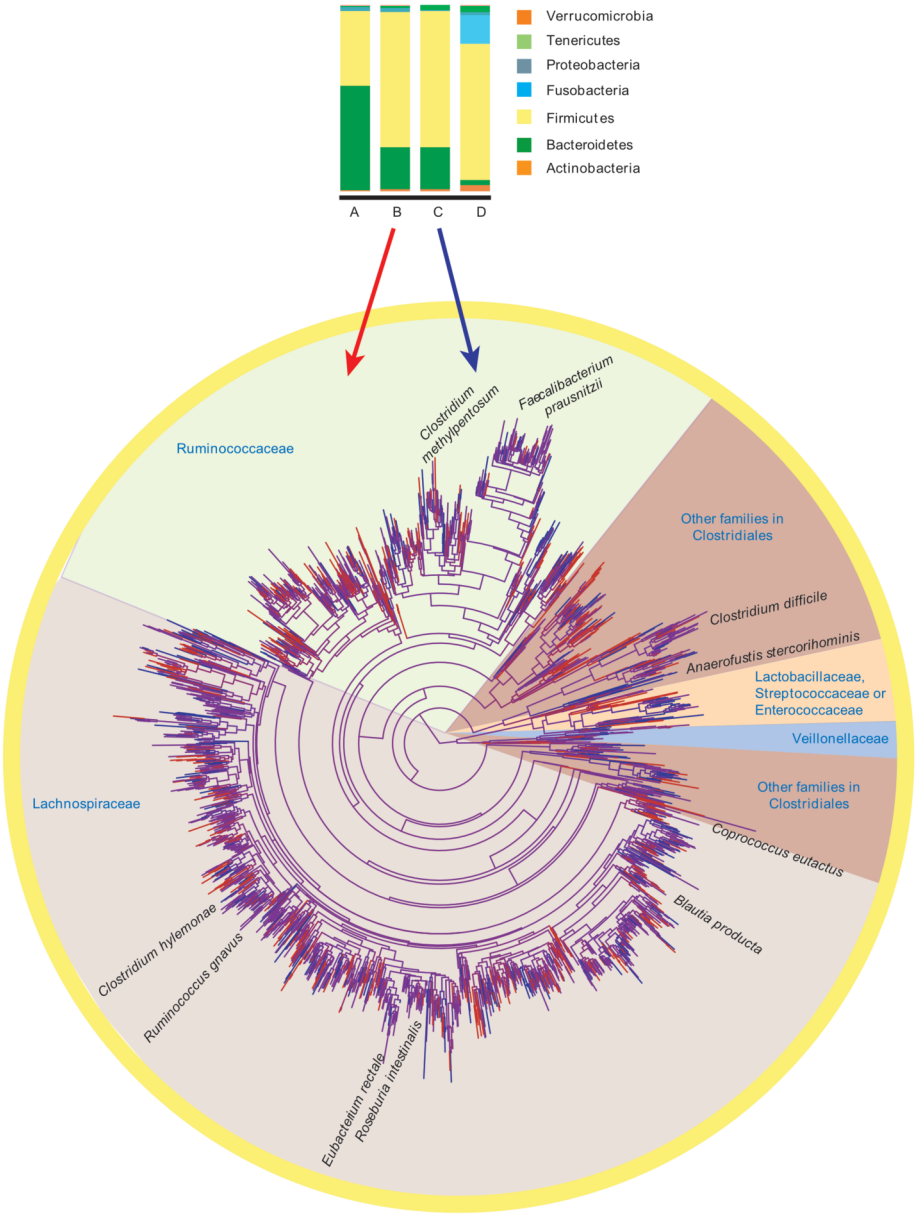


Figure 1.1: Demonstration of the taxonomic diversity of the gut microbiota and its variability between individuals. Top, a bar chart showing the variation in the taxonomic abundance profiles at the phylum level for four different individuals from the Unites States. Below, a phylogeny focused on the Firmicutes identified in individuals B and C coloured red if present in B, blue if present in C, and purple if in both. This shows the fine level variation between individuals who at higher levels appear taxonomically similar. Figure from a review by Lozupone *et al.*. (Lozupone et al. 2012b)

of UC patients, accompanied by relative increases and decreases in specific taxa (Michail et al. 2012). Although there is evidence that differences between UC patients and healthy controls are of a smaller magnitude than those observed with CD (Halfvarson et al. 2017; Willing et al. 2010).

Studies have also found significant differences in patients suffering irritable bowel syndrome (IBS). IBS refers to chronic discomfort of the bowels including cramping, abdominal pain, constipation and/or diarrhoea in the absence of the chronic inflammation that is characteristic of IBD. Studies have found both changes in diversity and taxonomic profiles of IBS patients (Rajilić-Stojanović et al. 2011; Pozuelo et al. 2015; Shankar et al. 2015). Furthermore, IBS diagnoses can be split symptomatically into sub-types, which can also be differentiated based on gut microbiota composition (Pozuelo et al. 2015; Kassinen et al. 2007).

Obesity

Initial observations of the association between gut microbiome composition and obesity were made using mouse models, under the hypothesis that microbiota might influence host energy storage by metabolising indigestible dietary intakes. Comparing lean and obese mice, obese mice were found to have increased abundances of Firmicutes and decreased levels of Bacteroidetes bacteria relative to lean mice (Ley et al. 2005). Whilst similar differences in the Firmicutes and Bacteroidetes ratio were reported by the same group in a human study (Ley et al. 2006), these findings have not replicated since (Sze and Schloss 2016). However, other taxonomic and functional differences have been observed with human obesity.

A study of 150 twins identified both taxonomic and functional differences distinguishing lean and obese individuals. The microbial functions differing between the groups were enriched for pathways related to carbohydrate and amino acid metabolism (Turnbaugh et al. 2009). A further study using over 1000 twins from the TwinsUK cohort, found that overall alpha diversity was reduced in the gut microbiome of obese individuals and identified an association with the genus *Christensenella*, which was at increased abundance in lean twins and conferred protection against weight gain when transplanted into mice (Goodrich et al. 2014b).

Type 2 Diabetes

Following observations that the gut microbiome associated with obesity, studies investigated if there were similar associations with type 2 diabetes (T2D) another metabolic linked disease. One of the key studies compared the gut microbiome, both its composition and function, between >70 individuals with T2D and >70 healthy controls from a Chinese pop-

ulation. Carrying out association analyses they identified bacterial species and functional pathways that were enriched in either the T2D or control groups and replicated several associations in an independent set of over 200 individuals. Amongst others, these included a decrease in butyrate production and *Faecalibacterium prausnitzii* abundance in the T2D patients, and differential amino acid and carbohydrate metabolism between the groups (Qin et al. 2012).

Other diseases

Differences in the taxonomic composition of the gut microbiota have been observed between colon cancer patients and healthy controls (Baxter et al. 2016). Beyond the GI tract, properties of the gut microbiome have also been associated with a diverse range of conditions including autoimmune diseases such as rheumatoid arthritis (Zhang et al. 2015) and with neurological conditions such as Parkinson's disease (Sampson et al. 2016) and potentially autism (Hsiao et al. 2013). Studies investigating ageing and frailty (a measure of healthy ageing) have also found differences in the gut microbiome associated with overall health and fitness (Claesson et al. 2012; Jeffery, Lynch, and O'Toole 2016).

There can be much variability in the approaches, both experimentally and analytically, used to assay gut microbiota (see Section 1.3). This can lead to poor reproducibility when comparing associations within the same disease (Sze and Schloss 2016). This also reduces the ability to compare associations between diseases to identify common gut microbiota effects. This is addressed in Chapter 6 where I carry out a comparison of 38 different health disorders with the same dataset.

Delineating causality

Whilst there have been some longitudinal and intervention studies focusing on the gut microbiota (Zeevi et al. 2015; Hjorth et al. 2017; Freedberg et al. 2015); human gut microbiome studies are, for the most part, observational. As such, they cannot delineate observed associations to determine if differences in the microbiome are cause or consequence of health effects. Most evidence for causal influences of the gut microbiota on health have come from experimental studies using model organisms.

In human research, interventions are limited to safe medications or lifestyle changes and by necessity cannot probe disease states directly. Studies using model organisms such as mice are able to control both gut microbiota composition and disease state. Gut microbiota can be manipulated through the use of antibiotic ablation or germ-free conditions to deplete the microbiota (Turnbaugh et al. 2008; Buffie et al. 2015). Subsequent exposure to specific taxa or microbiota transplantations can then be used to introduce desired taxa. Contrasting disease states can be probed by either inducing disease, such the experimental

colitis induced by dextran sodium sulphate exposure in mice, or by comparing comparable genetic models of human disease to healthy controls (Mizoguchi 2012). These approaches have been used extensively to probe causality of gut microbiota effects in disease. However, disease models are imperfect. The models might inaccurately represent human disease pathology and the microbiota of model organisms are different to those of the human gut.

The latter issue can be addressed by using models with humanised microbiota or by human microbiota transplants into germ-free models. These approaches provide evidence that the human gut microbiota can promote disease. For example, transplantation of gut microbiota from a Parkinson's disease patient to a mouse model of the disease exacerbates symptoms more than compared to using a healthy donor (Sampson et al. 2016). Germ-free models have also been used extensively to demonstrate a causal role for the gut microbiota in obesity (see Section 1.2.3).

Further evidence for a causal role of the gut microbiome in human health comes from human-human faecal microbiota transplants (FMT). This involves administration of homogenised faecal matter from a healthy donor directly to the gut of an afflicted patient. FMT has seen great success in the treatment of recurrent *Clostridium difficile* infection, often caused when *C.difficile* predominates in the gut following disruption of typical microbiota by antibiotic use (Leffler and Lamont 2015). It is thought FMT restores a patients gut microbiota preventing *C.difficile* becoming re-established (see Colonisation Resistance in Section 1.2.3) (Grehan et al. 2010). It has proved more successful than existing antibiotic based approaches to treatment (Brandt et al. 2012) and provides direct evidence that human gut microbiome composition can both cause and prevent disease.

1.2.3 Mechanisms underlying gut microbiota-disease associations

Regardless of whether differences in the gut microbiota are a response to or cause of disease, uncovering the mechanisms underlying host-microbiota interactions can improve our understanding of disease aetiology. Combining human association studies with experiments using animal models, several host-microbiota interaction mechanisms have been identified. Here, I discuss three major mechanisms (immune interactions, metabolic interactions, and colonisation resistance) and their relation to the diseases discussed in Section 1.2.2.

Immune Interactions

The GI tract is one of the largest interfaces between the immune system and micro-organisms. A range of innate and adaptive immune processes occur throughout the GI epithelial mucosa and in concentrations of lymphatic tissues called Payer's patches. Intestinal epithelial cells (IEC) promote innate immunity through the production of a mucus lining, which, along side the cells themselves, acts as a physical barrier separating the host and gut micro-

biota. IECs, in particular Paneth cells, also secrete antimicrobial peptides (AMPs) targeting specific taxa and can express receptors (called pattern recognition receptors, PRR) that detect specific antigens and stimulate both pro and anti-inflammatory responses. In the lower epithelium is the lamina propria, a layer of connective tissue that harbours both B and T-cells. These are exposed to antigen in the intestinal lumen and are involved in specific adaptive immune responses. The differentiated cells in the lamina propria migrate from Payer's patches, dense areas of lymphoid tissue distributed along the GI tract, within which B-cells and T-cells are specifically exposed to antigens from the intestinal lumen (Maloy and Powrie 2011; Shi et al. 2017; Round and Mazmanian 2009). These systems work together in conjunction with gut commensal microbes to maintain homoeostasis of both the host immune system and gut microbiota composition. Loss of this homoeostasis can lead to GI, and potentially systemic, inflammation and is one of the main mechanisms by which the gut microbiota influence host health (Round and Mazmanian 2009). Here, I provide a brief overview of some of the specific interactions that have been identified between the GI immune system and the gut microbiota and their potential roles in human disease.

Immune Development Contact with gut microbiota is required for normal development of GI immunity. Germ-free mice exhibit developmental defects in various aspects of the GI immune system including a reduction in lymphocytes within the intestinal epithelium, a reduction in antibody producing plasma cells in the lamina propria and less developed Payer's patches, as well displaying a range of morphological defects along the GI tract (Round and Mazmanian 2009). Changes in gut microbiota composition during the development of GI immunity in humans (see Age in Section 1.2.4) could be responsible for aberrant immune responses in later life.

Epithelial Barrier Integrity The segregation of the host and gut microbiota provided by the GI epithelial mucosa prevents chronic inflammatory responses to microbial antigens. Decreased mucus or AMP production by IECs can lead to invasion of the epithelial barrier by microbes and microbial antigens that promote inflammatory responses (Koslowski et al. 2010; Maloy and Powrie 2011). It has also been shown that gut microbiota can degrade mucin in the absence of other suitable substrates, which similarly compromises barrier function and promotes inflammation (Crost et al. 2013; Desai et al. 2016). The balance of mucus production and degradation is a key factor in maintenance of GI barrier integrity. The mucus layer can also influence the taxonomic composition of the gut microbiome. The AMPs released into the mucus by IECs act against specific microbes, whilst mucins can act as scaffolds for adhesion or metabolic substrates for commensal bacteria (Maloy and Powrie 2011). Several microbial strains have developed adaptations to utilise mucus glycoproteins (Crost et al. 2013; MacKenzie et al. 2010), which gives them a competitive colonisation advantage over other taxa. The balance of these host and microbial selective processes for mucosa adhesion and growth may be key for maintaining a beneficial, or

non-detrimental, gut microbiota composition.

Beyond the mucus layer, gut microbiota can also influence the integrity of the IECs themselves. Commensal antigen sensing by PRRs (such as Toll-like receptors) can promote proliferation of IECs (Hsu et al. 2010), prompt AMP production (Vaishnava et al. 2008), and strengthen the inter-cellular tight junctions that prevent microbial invasion between IECs (Maloy and Powrie 2011). Commensal microbiota are therefore required to maintain a robust IEC barrier against potential pathogens.

Immune Stimulation Gut microbiota and their metabolic products can directly stimulate both pro and anti-inflammatory responses in the GI immune system. Highly conserved antigens found across a range of microbes are referred to as pathogen-associated molecular patterns (PAMPs). PAMPs include bacterial cell membrane components such as peptidoglycans and, in the case of gram-negative bacteria, lipopolysaccharide (LPS). These are bound by PRRs on IECs and other immune cell receptors within the epithelial mucosa (Hakansson and Molin 2011). PAMPs stimulate a number of pro-inflammatory responses including induction of cytokines, such as tumour necrosis factor-α and interleukin-6, which can promote both innate and adaptive immune cascades (Maloy and Powrie 2011). Increased levels of taxa expressing PAMPs can promote chronic GI inflammation. Epithelial invasion by LPS in the gut can also lead to systemic inflammation termed endotoxemia, which has been linked with development of T2D (Cani et al. 2007; Cani et al. 2008).

Gut microbiota can also produce anti-inflammatory metabolites such as short-chain fatty acids (SCFAs). SCFAs (butyrate, propionate and acetate) are produced by microbial fermentation of non-digestible carbohydrates from the host diet and are one of the main microbiota derived energy sources. SCFAs also have anti-inflammatory properties and, as discussed in Section 1.2.2, a reduced abundance of SCFA producing taxa has been observed in IBD and T2D patients (Sokol et al. 2008). SCFAs stimulate IECs and GI immune cells by binding SCFA specific receptors, such as the G protein coupled receptors GPR43 and GPR41, resulting in a range of anti-inflammatory effects (Shi et al. 2017; Maslowski et al. 2009). These include an increased proliferation of GI T_{reg} cells (a class of immune cells which modulate pro-inflammatory responses of other T cells and are important in prevention of autoimmunity) (Smith et al. 2013), a moderation of pro-inflammatory responses to bacterial LPS (Vinolo et al. 2011), and stimulation of AMP production by IECs (Zhao et al. 2018). SCFAs can also promote IEC proliferation, which may aid maintenance of epithelial barrier integrity (Park et al. 2016). Interestingly, butyrate can also promote apoptosis and have a protective effect against colon cancer. This is due to a high concentration of butyrate in cancerous cells which preferentially metabolise glucose. This leads to butyrate inhibition of histone deacetylase enzymes and cell death (Donohoe et al. 2012). There are also examples of specific anti-inflammatory molecules produced by individual commensal species. The first identified was bacterial polysaccharide A (PSA) produced by *Bacteroides fragilis*.

Introduction of PSA producing *B.fragilis* in the GI tract is protective against experimental induction of colitis in mice, whereas PSA deficient mutants are not (Mazmanian et al. 2005). PSA, similar to SCFAs, induces proliferation of T_{reg} cells that moderate GI inflammation (Mazmanian, Round, and Kasper 2008).

It is believed that the balance of pro and anti-inflammatory taxa in the gut microbiota could determine the state of GI inflammation, which, in turn, could influence the ability of different commensals to colonise the epithelium. Diseases of chronic GI inflammation, such as Crohn's disease and ulcerative colitis, might result from extreme disequilibrium of this system. However, it is not clear if this is initiated by inflammatory changes, for instance due to host genetic variations, or changes in gut microbiota composition, for instance after pathogen invasion or antibiotic treatment.

Metabolic Interactions

The combined genomic complement of the gut microbiota encodes a diverse range of enzymes from various metabolic pathways, including many not found in the human genome. The GI tract is exposed to host dietary intakes and various metabolites produced and excreted by the host. Gut microbiota metabolise these into numerous other metabolites that can directly affect the host, such as in the immune examples described previously. Here I summarise a selection of other major metabolic actions of the human gut microbiota that can influence host health.

Energy Uptake The gut microbiota can metabolise a range of different substrates, including protein, fats, and carbohydrates, to small molecules that increase the bioavailability of dietary energy to the host. This is most simply demonstrated by mice reared in germ-free conditions, which gain less weight than conventionally reared mice fed a similar diet (Bäckhed et al. 2007). As described in Section 1.2.2, differences in gut microbiota composition are also observed between obese and lean humans. These differences are correlated with differences in diet, however there is also evidence that differential microbiota composition can be a causal factor in obesity. Germ-free mice inoculated with microbiota from obese human donors gain weight more readily than those receiving microbiota from lean donors (Goodrich et al. 2014b; Ridaura et al. 2013). This is likely due to increased energy extraction by the obese gut microbiome, which contains a higher abundance of genes associated with carbohydrate metabolism (Turnbaugh et al. 2006). Further evidence that metabolism of the human gut microbiota could have a causal role in obesity comes from Roux-en-Y (RNY) gastric bypass surgery, a treatment for obesity where the GI tract is altered to bypass most of the stomach and duodenum. The procedure has a rapid and profound effect on metabolic markers that occur faster than could be expected due to dietary changes. RNY bypass surgery is also associated with rapid and large shifts in gut microbiota composition

that could be responsible for observed benefits (Tremaroli et al. 2015; Li et al. 2011). The importance of the effects of human gut microbiota on energy absorption was also highlighted by a large personalised diet study in Israel. This profiled gut microbiome composition and monitored glycemic responses after different dietary intakes in 800 individuals. It found that individuals had very different glycemic responses to the same foodstuffs. However, an individual's glycemic responses could be predicted from the taxonomic composition of their gut microbiome (Zeevi et al. 2015). The metabolic potential of the gut microbiota is therefore an important influence on host energy uptake from diet and could have important roles in the development of obesity, T2D and metabolic syndrome.

Bile Acids Bile acts as a detergent to improve the solubility and digestion of lipids in the small intestine and predominantly consists of bile acids. Primary bile acids (chenodeoxycholic acid, CDCA and cholic acid, CA) are produced by the liver where they are conjugated to glycine or taurine. These then accumulate in the gall bladder before being secreted into the GI tract. Here, specific taxa in the gut microbiota are able to de-conjugate and carry out 7α-dehydroxylation of primary bile acids to produce secondary bile acids (lithocholic acid, LCA and deoxycholic acid, DCA). Both secondary and primary bile acids are absorbed in the lower GI tract and recycled via the liver in a loop that feedbacks and regulates further bile acid production (Ridlon et al. 2014; Staley et al. 2017). Several studies suggest that differences in the the gut microbiota observed with disease may result from bile acid mediated effects.

It has been shown that metabolism of primary bile acids to secondary by the gut microbiota has an agonistic effect on the farnesoid X receptor (FXR), which regulates the production of bile acids (Sayin et al. 2013). As a result, germ-free mice display higher levels of bile acid relative to conventionally raised mice. Differences in bile acid levels have been observed in several diseases (Kakiyama et al. 2013; Wildenberg and Brink 2011; Tomkin and Owens 2016; Mudaliar et al. 2013), and disruption of bile acid production may be one mechanism by which gut microbiota influence host health. For example, there is a lower abundance of bile acid metabolising taxa in the gut microbiota of liver cirrhosis patients compared to healthy controls (Kakiyama et al. 2013). This coincides with a reduced total bile acid pool and a relatively lower level of secondary bile acids. Activation of FXR has also been shown to improve insulin sensitivity, suggesting gut microbiota could also influence metabolic syndrome via bile acid mediated mechanisms (Mudaliar et al. 2013). However, bile acid homoeostasis is maintained by a feedback loop and primary bile acid production inevitably proceeds conversion to secondary bile acids. Hence differences in gut microbiota composition might also reflect responses to changes in bile acid production.

Bile acids have a direct influence on gut microbiome composition. Large phylum level shifts in gut microbiota composition were observed with administration of dietary CA in rats. There was a loss of Bacteroidetes with an increased abundance of Firmicutes including

Clostridia, which is known to contain species able to carry out 7α-dehydroxylation of primary bile acids (Islam et al. 2011). This likely reflects a competitive advantage for these taxa leading to an increased proliferation relative to other gut microbiota. However, it has also been proposed that the increased hydrophobicity of DCA relative to CA could also have direct effects in disrupting cell membranes and actively suppress growth of susceptible taxa (Ridlon et al. 2014).

There is also evidence of a specific association between bile acids and *C.difficle* infection. Whilst GI *C.difficile* invasion following antibiotics is typically considered as the pathogen capitalising on reduced competition (see Section 1.2.2 and Colonisation Resistance), antibiotic use also results in a reduction in primary to secondary bile acid conversion (Theriot, Bowman, and Young 2016). Primary bile acids can prompt germination of *C.difficile* spores and may, in part, be responsible for the increased risk of infection. This is supported by evidence that germination of *C.difficile in vitro* is promoted or inhibited when applying concentrations of bile acids observed in patients before and after FMT respectively (Weingarden et al. 2016).

It is not clear if bile acid changes under disease are predominantly driven by host or gut microbiota effects, however bile acid metabolism could be an important mechanism by which the gut microbiota influence host health.

Other Metabolic Processes Gut microbiota contribute a diverse range of other metabolic processes that could influence host health. Methane production by methanogens has been linked to GI transit time, which is in turn associated with colon cancer (Triantafyllou, Chang, and Pimentel 2014). Metabolic syndrome and T2D are associated with an increased level of circulatory amino acids (Neis, Dejong, and Rensen 2015). Gut microbiota contribute to the systemic amino acid pool, both via proteolysis and synthesis, and can influence host amino acid metabolism (Mardinoglu et al. 2015; Neis, Dejong, and Rensen 2015). Thus, amino acid metabolism could be a further mechanism by which gut microbiota contribute to metabolic disease. Gut microbiota metabolism can also influence disease treatments by metabolising medications. For example, gut microbiota composition can moderate the efficacy and toxicity of chemotherapy (Alexander et al. 2017). Overall, there are many different ways in which the metabolic action of the gut microbiome could influence host health and further research is required to determine which processes have the largest influence on individual diseases.

Colonisation Resistance

The human gut microbiota protect against colonisation by enteric pathogens. Disruption of gut microbiome communities by antibiotic use can lead to an increased risk of infection by bacterial pathogens such as *C.difficile*, pathogenic *Escherichia coli* and *Klebsiella pneu-*

moniae, particularly when a pathogen has developed resistance to the antibiotic used (Zerr et al. 2016; Leffler and Lamont 2015). FMT, which reintroduces commensal communities to the colon, can restore colonisation resistance providing evidence that the susceptibility to infection could be due to a loss of commensal gut microbiota (see Section 1.2.2). Several of the host-microbiome interactions discussed previously serve to promote colonisation resistance including stimulation of AMP production, promotion of IEC barrier integrity, and regulation of microbial spore germination by secondary bile acids. However, host independent inter-microbial effects can also drive colonisation resistance. These can be either passive or active interactions.

Passive mechanisms of resistance are driven by non-pathogenic taxa competing with pathogens for environmental resources. One study demonstrating this effect used the model of *Citrobacter rodentium* infection in germ-free mice. These were found to be more susceptible to enteric infection than conventionally raised mice. Reintroduction of specific commensal microbes after pathogen colonisation caused dramatic reductions in the relative levels of GI *C.rodentium*. By repeating this experiment with either mono or polysaccharides as dietary carbohydrates and inoculating *C.rodentium* infected mice with commensals specialised to either of these dietary types, it was shown that this effect is mediated by competition for energy substrates. Commensal displacement of pathogen colonisation only occurred when the commensals were able to metabolise the carbohydrates and compete with the pathogen for resources (Kamada et al. 2012). As discussed previously, gut microbiota communities contain a diverse range of metabolic capabilities and hence will compete for resources with, and limit colonisation of, a range of pathogenic bacteria.

Bacteria can also actively target other species to inhibit their growth. Bacteriocins are anti-microbial peptides that are secreted by bacteria to moderate the growth of competing species whilst having no effect on the producer. Bacteriocin genes are widespread in genomes of the human gut microbiota (Donia et al. 2014). Commensals producing pathogen-specific bacteriocins actively contribute to colonisation resistance. For example, the commensal *Bacillus thuringiensis* produces bacteriocins targeting *C.difficle* (Rea et al. 2010); *Lactobacillus salivarius* produces a bacteriocin targeting *Listeria monocytogenes*, a foodborne pathogen that can cause Listeriosis; and a strain of *Enterococcus faecalis* produces a bacteriocin targeting Vancomycin-resistant enterococci (Kommineni et al. 2015), which highlights the potential of microbiome communities as sources of novel antimicrobial agents to overcome increasing levels of resistance to current antibiotics.

1.2.4 Host determinants of gut microbiome composition

The gut microbiome can influence a range of diseases via numerous potential mechanisms. However, the host is also a major influence on the gut microbiome and its composition is associated with a range of host environmental and lifestyle factors.

Age

Differences in gut microbiome composition with age have been observed in several cross-sectional and longitudinal studies (Yatsunenko et al. 2012; Odamaki et al. 2016). Although recent evidence in animal models suggests that prenatal microbial transmission from mother to foetus could occur (Moles et al. 2013; Jiménez et al. 2008), the uterine environment is typically considered sterile and from birth the gut microbiome starts in a low diversity state (Odamaki et al. 2016; Yatsunenko et al. 2012). Infant gut microbiota are then acquired via several mechanisms including exposure to maternal vaginal and faecal microbiota during birth, microbial exposure during feeding, and other environmental exposures (Koenig et al. 2011). Initially, the neonatal gut microbiota is enriched for Actinobacteria able to metabolise lactate, with most belonging to the genus *Bifidobacterium* (Koenig et al. 2011; Odamaki et al. 2016). As the relative contribution of milk to an infants diet decreases there is a concurrent change in microbiome composition, with a relative increase in other taxa able to degrade complex foods such as plant polysaccharides (Koenig et al. 2011). The rapid accumulation of novel taxa from environmental exposures and changing dietary habits leads to a period of instability in the gut microbiota that lasts until around three years old, where the gut microbiome forms a more stable adult-like composition (Koenig et al. 2011; Yatsunenko et al. 2012; Odamaki et al. 2016). It has been proposed that compositional changes due to exposures in the unstable infant period, such as differences in mode of delivery or antibiotic use, could have potentially long-term impacts on the gut microbiota and might influence host immunity and health in later life (Bäckhed et al. 2015; O'Neill, Schofield, and Hall 2017). After progressing to adulthood the configuration of the gut microbiota remains stable with communities largely dominated by Bacteroidetes and Firmicutes. In the extremes of ageing there is then a marked decrease in taxonomic diversity and increases in the relative abundances of the genera *Eubacterium* and *Bacteroides* (Odamaki et al. 2016). This age associated dysbiosis has been shown to coincide with increased frailty and differences in dietary habits and environmental exposures in the elderly (Claesson et al. 2012).

Geography

The majority of cohort-based studies aiming to profile typical gut microbiota composition have focused on Western populations from the USA and Europe (Consortium 2012; Falony et al. 2016; Zhernakova et al. 2016; Qin et al. 2010). However, several smaller studies have also profiled 'non-Western' populations. Yatsunenko *et al.* examined gut microbiome profiles of individuals across a range of ages from Malawi and Venezuela and contrasted them against similar individuals from the USA. They observed a notable taxonomic divergence in the Malawian and Venezuelan samples compared to those from the USA (Yatsunenko et al. 2012). Differences between Western and non-Western populations were also observed

in the Flemish Gut study. When comparing Belgian gut microbiome profiles to cohorts from the UK, USA, and the Netherlands there was considerable taxonomic overlap, with 17 core genera observed that accounted for a median of 72% of gut microbiota across these individuals (Falony et al. 2016). However, expanding comparisons to include individuals from Papua New Guinea, Peru, and Tanzania reduced the number of core genera to 14, due to their increased divergence from the Western populations. More extreme taxonomic differences have also been observed when comparing Western gut microbiota profiles to those of hunter-gather groups such as the Hadza people from Tanzania, whose lifestyles are often considered as more representative of those experienced during human evolution (Rampelli et al. 2015).

Across all geographical comparisons there are consistent trends (Gupta, Paul, and Dutta 2017). Hunter-gather tribes have a highly diverse microbiota, often characterised by a dominance of the genus *Prevotella*. Whereas Western populations in general have a relatively less diverse gut microbiota enriched for taxa including *Bacteroides* and *Bifidobacterium*. The gut microbiota of rural populations displays properties in-between the two and might represent the transition between these extremes of lifestyle. The geographic distribution of populations is inherently associated with differences in host genetics and cultural factors such as diet. Whilst it is not yet clear which are the major drivers of the observed differences in the gut microbiome, it is likely that both geographic separation and cultural differences contribute; as differences have also been observed between comparable populations from different countries (Li et al. 2014) and between ethnic groups within the same geographic area (Consortium 2012). The existence of geographically distinct gut microbiota profiles necessitates replication of gut microbiome studies in different populations to confirm the robustness and applicability of observations. This is considered in Chapter 7, where I compare the similarity of microbial co-occurrence networks across three geographically diverse cohorts.

Diet

As discussed in Section 1.2.3, the gut microbiome acts as a direct interface between dietary intakes and host metabolism with consequences for host health. However, as might be expected, several studies have also shown that dietary changes can also cause large-scale changes in gut microbiota composition. These have shown that diet-associated differences in taxonomic abundances in the gut reflect taxa abilities to metabolise predominant dietary components. For example, *Bacteroides* dominant gut microbiomes have been associated with high protein and fat rich diets, whereas *Prevotella* dominance is associated with higher carbohydrate and fibre intake (Wu et al. 2011). These associations are not limited to long term dietary behaviours, for example swapping between meat and plant based diets can cause rapid changes in the levels of bacteria known to metabolise plant fibres (David et al. 2014). The ability to alter microbiota composition with dietary interventions is a promising

avenue of future research. For instance, it could be possible to identify dietary components that promote taxonomic compositions that have positive influences on host immunity or promote resilience to pathogen colonisation (see Section 1.2.3). This has been the aim of several pre-biotic intervention studies, which aim to provide specific dietary substrates that promote the growth of targeted beneficial bacteria in the gut (Beserra et al. 2015; Ford et al. 2014).

Genetics

One of the first large-scale studies to investigate general associations between host genetics and the composition of the gut microbiota was carried out in the TwinsUK cohort. This utilised its twin structure to estimate the heritability of individual taxonomic abundances (Goodrich et al. 2014b). The study identified significant genetic influences on a range of taxa, although heritability estimates were for the most part low. The most heritable taxa was the family Christensenellaceae, which had roughly 40% of the variance in its abundance attributable to additive genetic effects. There have also been several studies that have demonstrated differences in the gut microbiome associated with specific genomic variants of interest (Frank et al. 2011; Rausch et al. 2011), and more recently some significant associations with gut microbiome traits have been described in genome wide associations studies (Bonder et al. 2016). For example, the abundance of *Bifidobacterium*, a genus of lactose metabolising taxa, is associated with variants in the lactase gene locus, which determines the host's ability to digest lactose (Blekhman et al. 2015). In this case, microbial changes likely reflect genetic driven dietary habits, namely reduced milk consumption. However, for the most part, the mechanisms mediating genetic influences on the gut microbiome are yet to be determined.

Health and Medication

The gut microbiome has been shown to associate with health. However, the directionality of many of these effects is unknown and, as evidenced by the mechanisms discussed in Section 1.2.3, could also be bidirectional. Thus differences in gut microbiome could also represent a response to changes in host health. It is also necessary to take into account the influence of disease treatment on the gut microbiome. As might be expected, oral antibiotics have a rapid and wide-scale effect on the taxonomic composition of gut microbiota reducing overall diversity (Jakobsson et al. 2010). In some cases these effects can be persistent and detected in the gut microbiome years later (Jernberg et al. 2007). The anti-diabetic drug metformin has also been shown to influence gut microbiome composition and function (Forslund et al. 2015). This confounded initial results from diabetes-gut microbiome studies and highlights the importance of considering medication use. I examine the influence of common medications on the gut microbiota in Chapters 4 and 6.

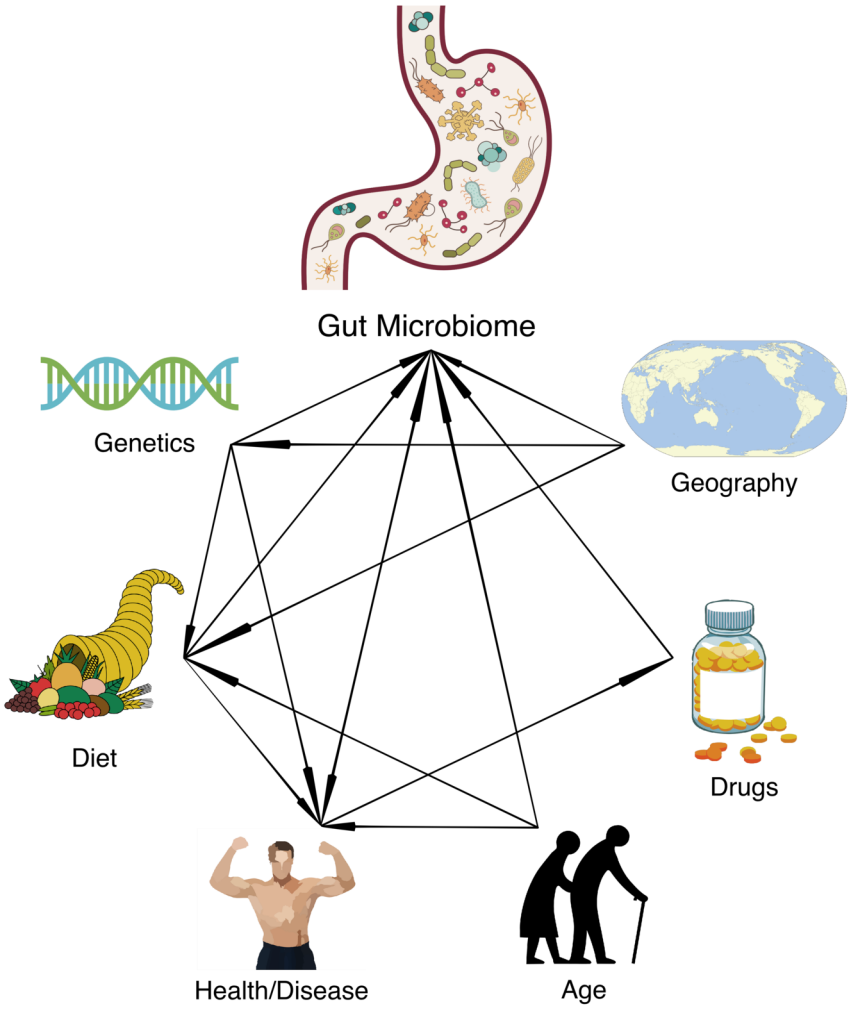


Figure 1.2: Connectivity of influences on the human gut microbiome.

Not only do all the aforementioned factors influence the gut microbiome, they can also interact with one another, and, in the case of factors such as health, could also be influenced by the gut microbiome itself (Figure 1.2). This leads to a complex and dynamic system that can make it challenging to studying the role of the gut microbiome in host health. The majority of the analyses reported in this thesis utilise samples from twin pairs. This provides a means to control for at least some of the genetic and environmental exposures shared between them.

1.3 16S rRNA gene profiling of microbiota

1.3.1 Phylogenetics and the ribosome

Phylogenetics is concerned with describing the evolutionary relationships between organisms. This is principally achieved through comparison of gene sequences. By isolating a region of gene sequence and aligning the comparable parts between different organisms, it

is possible to observe the mutations unique and shared between individuals and infer the evolutionary pathways leading to the observed sequences. This allows quantification of the evolutionary similarity, or distance, between organisms and creation of phylogenetic trees describing these relationships. Phylogenetic analyses require a gene that is found in the genomes of all the organisms under investigation, most typically these are ribosomal RNA genes (Weisburg et al. 1991).

Ribosomes are responsible for catalysing the translation of messenger RNA sequences to peptides; a highly conserved process occurring in all living cells. These intracellular structures consists of subunits made from a combination of proteins and non-coding rRNAs. In prokaryotes, the smallest ribosomal subunit contains the 16S rRNA. The importance of its function has lead to high evolutionary conservation (low mutation rates) in the functional regions of the gene encoding the 16S rRNA. However, these functional regions are interlinked by spacer regions, termed variable regions, that are under less selection pressure and more readily accumulate mutations (Figure 1.3). These properties provide a powerful tool for phylogenetic analyses. The known sequence in the conserved regions can be used to align gene sequences and also facilitates the design of universal PCR primers to assay 16S rRNA genes from different organisms. Whilst the variable regions in-between make it possible to make evolutionary inferences (Weisburg et al. 1991). These features have led the 16S rRNA gene to become a standard for prokaryotic phylogenetics and resulted in the production of large reference databases of 16S rRNA gene sequences from thousands of prokaryotic species. These have proved invaluable for profiling microbiota (DeSantis et al. 2006; Quast et al. 2012; Cole et al. 2008).

1.3.2 Sequencing 16S rRNA genes in the gut microbiome

Sequencing of 16S rRNA genes is one of the most widely used methods to generate taxonomic profiles of gut microbiota. Prior to the advent of modern sequencing technologies, 16S rRNA gene sequences were sequenced on an individual basis and microbiomes were assayed using culture based approaches that under-represented taxa not amenable to culturing. The development of high-throughput technologies, such as Roche 454 pyrosequencing and the Illumina dye based approach, has enabled sequencing of thousands of 16S rRNA gene sequences from microbiome samples (Sogin et al. 2006). These can then be mapped to databases of 16S rRNA gene sequences to identify the prokaryotes present in the community and their relative abundances. In practical terms this is a more complex process both in terms of sequencing the 16S rRNA genes and analysis of the read data.

The use of variable regions

The benefit of high-throughput sequencing technologies is the ability to sequence hundreds of thousands of reads providing deep sampling from a pool of sequences, however this

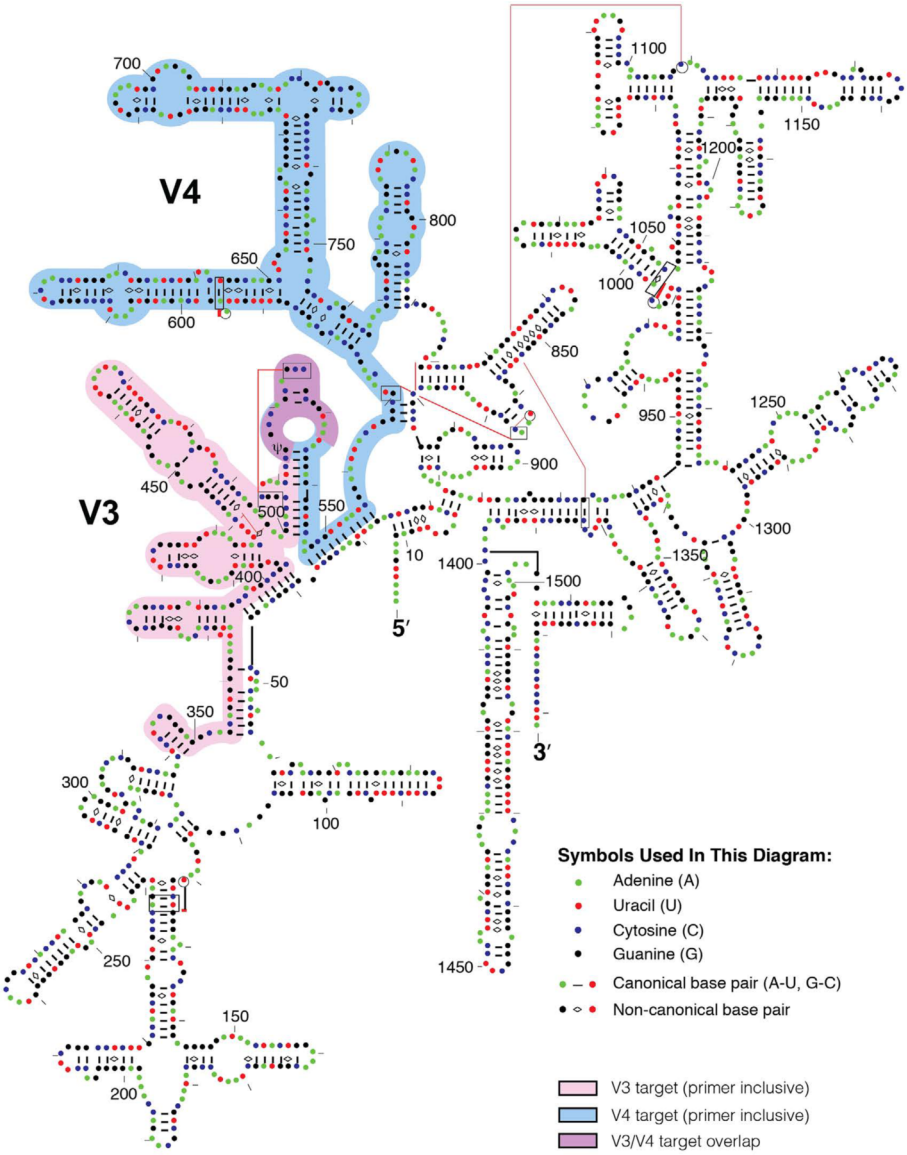


Figure 1.3: Diagrammatic representation of the secondary structure of an *Escherichia coli* 16S rRNA, highlighting the V3 and V4 variable regions. Adapted from Ziesemer *et al.* 2015 (Ziesemer et al. 2015).

comes at the cost of read length - pyrosequencing reads average around 400bp in length and Illumina sequencing can reach lengths up to 250bp (Goodwin, McPherson, and McCombie 2016). These limits mean that it is not possible to assay the whole 16S rRNA gene ($\sim$ 1.5kb in length (Brosius et al. 1978)). This has been addressed by only using select variable regions in 16S rRNA gene-based microbiota assays. The 16S rRNA gene contains nine variable regions and several have been used in microbiome studies. I will focus on the V4 region (Figure 1.3) and sequencing using Illumina platforms as this is used throughout this thesis. Illumina short-read sequencing has also become widely adopted in microbiome studies using 16S rRNA gene sequencing as it provides deeper sampling (more reads) than pyrosequencing (particularly important when multiplexing samples - Figure 1.4). This has in turn promoted the use of the V4 region as it is $\sim$ 253bp long and can be fully sequenced, with sufficient overlap, by Illumina 250bp paired-end sequencing.

DNA extraction and PCR

The first stage of sequencing 16S rRNA genes in a gut microbiome study involves extraction of microbiome DNA, typically from a faecal swab or sample. This can be carried out using a range of commercial kits using chemical and/or physical lysis steps to break open microbial cells. A variable region of the 16S rRNA gene is then amplified using PCR with universal prokaryotic primers targeting conserved gene regions. After PCR, amplicons are then sequenced producing raw 16S rRNA gene reads. PCR polymerase, primers, DNA extraction method, and sequencing platform can all influence the resulting microbiota profiles (Cruaud et al. 2014; Tremblay et al. 2015; Gohl et al. 2016; Walker et al. 2015). Standardised protocols have been produced that aim to reduce the experimental variance between studies, for instance the Earth Microbiome Project that developed the widely used 515F-806R primers for the 16S rRNA V4 region (Gilbert, Jansson, and Knight 2014). However, there is still much variability between the experimental steps between different studies that impedes replication of results across datasets. In Chapter 6, I carry out association analyses between multiple diseases within the same data set allowing a uniform comparison of the results.

Quality control

Quality scores from the sequencing platform can be used to filter low quality reads from a microbiome sample. An additional step to improve 16S rRNA marker gene sequencing is filtering of chimeric reads. Incomplete extension between primers during PCR can produce partial reads that can act as primers in the next round of amplification. These can bind to sequences from a different species generating reads that contain fragments from different organisms termed chimeras (Qiu et al. 2001). Several algorithms are available to identify and remove these sequences in 16S rRNA gene amplicon data. These use either reference

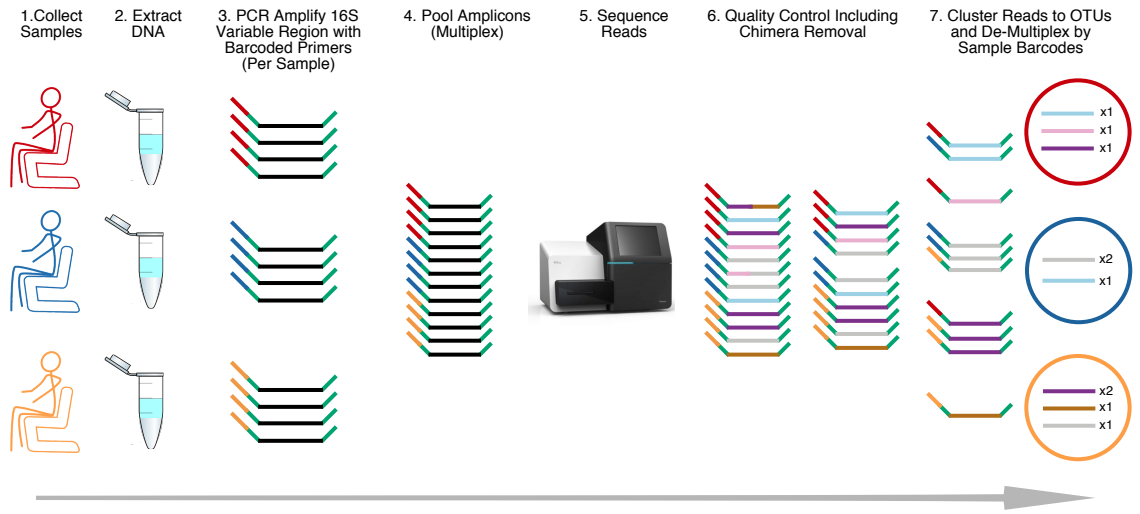


Figure 1.4: Overview of an experimental pipeline to generate 16S rRNA gene sequencing-based gut microbiota profiles at a large population cohort scale, highlighting the use of barcoded PCR primers. This enables pooling, or multiplexing, of reads so they can be sequenced within the same sequencing run, reducing costs and increasing efficiency. Reads can then be split later by sample based on the unique barcodes on each read.

sequences to identify chimeras from combinations of known sequences, or detect chimeras based on the abundance of other reads within the data (assuming a chimera will be at lower abundance than its source reads) (Edgar et al. 2011; Haas et al. 2011). Once the 16S rRNA gene read data in a sample has been cleaned it can be combined with reads from other samples in the study. An overview of the complete pathway from samples to 16S rRNA gene reads in a large cohort study is shown in Figure 1.4. Once the data are cleaned and combined the final stage before analysis of the microbiota profiles is the generation of operational taxonomic units.

1.3.3 Operational Taxonomic Units

Limitations of raw 16S rRNA gene sequences

The 16S rRNA gene reads produced from microbiome sequencing can contain many hundreds of thousands of reads for a given sample. As a result of the PCR amplification, the existence of multiple bacteria of the same type in a microbiome, and multiple copies of the 16S rRNA gene within bacteria (Větrovský and Baldrian 2013), there will be many duplicates. However, even after dereplication of reads there will be thousands of different unique sequences. This is due to:

1. The large diversity of hundreds to thousands of organisms in the gut microbiome with unique sequences across the 16S rRNA gene V4 region
2. Sequencing errors inducing small nucleotide changes producing new erroneous reads
3. PCR errors such as chimera formation creating new erroneous reads

Ideally, the influence of the last two factors would be minimal given sufficient quality control but current methods cannot account for all of these effects perfectly. The first factor is not problematic in terms of biological representation of microbiota but presents analytical challenges. The high dimensionality of these data sets leads to high computational demands for processing and storage space, increases the burden of multiple testing in statistical analyses, and can make them unsuitable for many statistical approaches due to there extreme sparsity (numerous reads are found at low levels in only one or two samples in a study).

There are also limits to the resolution achievable using sequence reads from the V4 region of 16S rRNA genes. It is only possible to assess the evolutionary differences within this limited region. These may not be representative of wider divergence across the whole 16S rRNA gene and even less so across the whole bacterial genomes (Kim et al. 2014). The nucleotide differences within the V4 region might over or underestimate the wider evolutionary divergence and don't directly represent functional differences between taxa (Schloss 2010). This is further confounded by bacterial exchange of genetic material via non-genomic vectors in horizontal gene transfer (Huddleston 2014). Mutation rates can also differ between variable regions and between microbial clades (Kim, Morrison, and Yu 2011; Yang, Wang, and Qian 2016). As a result of these factors, even if suitable analytical methods existed, it would not necessarily be beneficial to analyse 16S rRNA gene sequences at the unique read level. This is typically overcome by clustering reads together to operational taxonomic units.

The term operational taxonomic unit (OTU) was defined by ecologists to refer to groups of taxonomically related organisms that are grouped together to simplify quantitative analyses (Sneath and Sokal 1973). An OTU does not translate to any strict biological representation but is a unit simply used for analytical (or operational) purposes. Grouping to OTUs could be done at any taxonomic level based on genetic similarity or even based on non-genetic relationships such as morphology. However for the purposes of microbiota analyses, the OTU has taken on a more standardised definition.

In studies using 16S rRNA gene sequencing, OTUs are used to reduce the dimensionality of the read data. The unique reads observed across all samples in a study are clustered together into OTUs based on their sequence similarity. Read counts are then summed into their assigned OTUs to produce OTU tables, giving the counts of each OTU in each sample. A visual representation of OTUs is shown in Figure 1.5.

The 97% similarity threshold

To generate OTUs, the unique 16S rRNA gene reads across all samples in a study are clustered by sequence similarity to create groups of reads more similar than a given threshold. This is typically carried out at 97% - that is, grouping two reads into the same OTU if, when aligned, they share at least 97% of the nucleotides along their length. The 97% thresh-

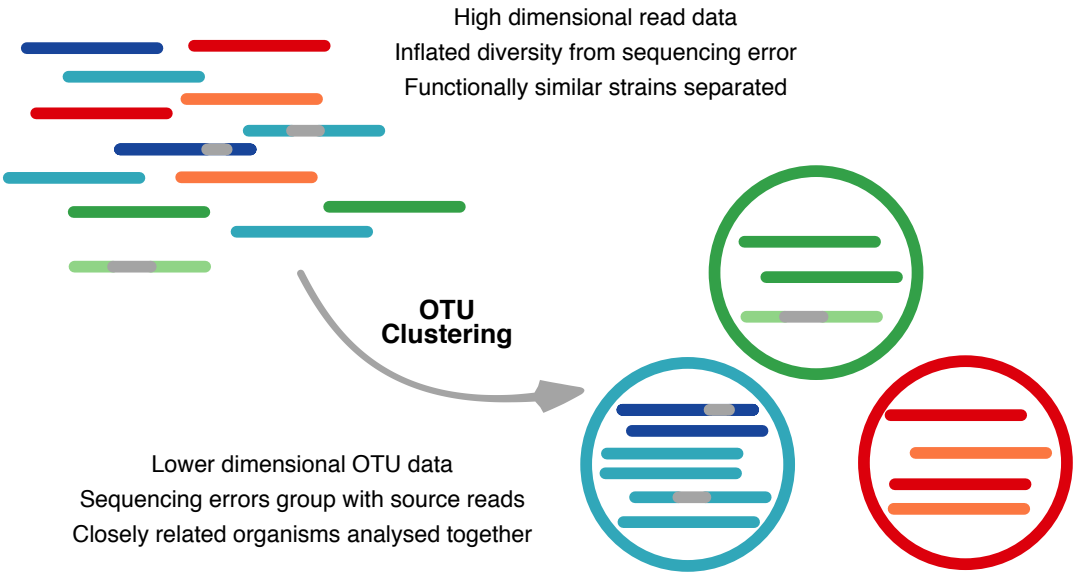


Figure 1.5: Schematic of the idealised concept of OTU clustering. Colours represent different source organisms, different shades indicate similar strains, and grey indicates sequencing errors.

old was selected as a result of studies comparing the genomic variance between bacterial species in relation to the variance in their 16S rRNA genes (Stackebrandt and Goebel 1994; Konstantinidis and Tiedje 2005). These found that at the 97% level the variance in the 16S rRNA genes was roughly equivalent to existing definitions of bacterial species based on genomic divergence. This threshold is also sufficiently flexible to group reads produced from sequencing errors with their source sequences (Huse et al. 2010). However, these studies considered variation across the full length 16S rRNA gene and this threshold is still applied when only considering much shorter variable regions. There is also evidence that different similarity thresholds might be required for different prokaryotic clades due to differences in mutation rates (Mysara et al. 2017). I carry out a comparative analysis to identify an optimal similarity threshold value in a comparison of OTU clustering approaches in Chapter 5.

Clustering 16S rRNA gene amplicon reads to OTUs

Most simply, clustering 16S rRNA gene amplicons to OTUs is an iterative process of carrying out pairwise alignments and comparisons and grouping reads together where they are sufficiently similar. However, in practice thousands of reads make this infeasible, requiring many millions of pairwise comparisons. This is overcome in two ways:

1. Using greedy algorithms that optimise comparisons between sequences to those likely to match (Edgar 2010; Rognes et al. 2016; Li and Godzik 2006). An example of which is shown in Figure 1.6.

2. Limiting comparisons by clustering experimental data to a restricted reference list of sequences (Navas-Molina et al. 2013; Schloss and Westcott 2011).

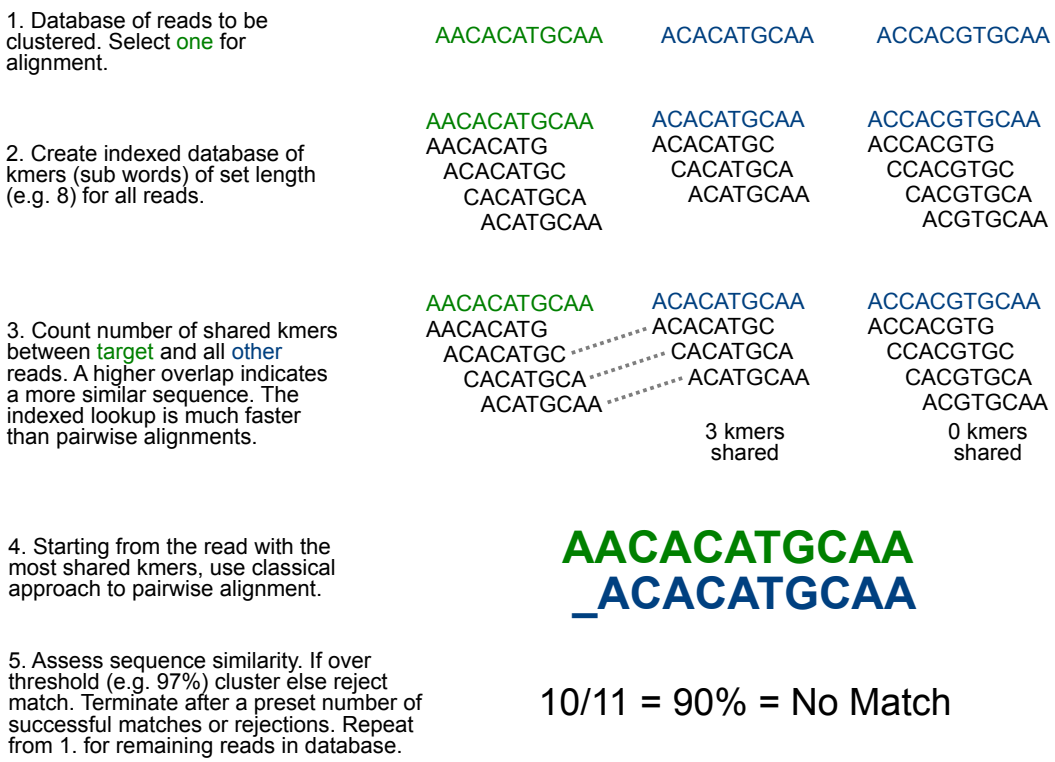


Figure 1.6: Overview of a greedy approach to clustering read data based on the kmer indexing approach used by VSEARCH (Rognes et al. 2016).

Both of these approaches have been used alone and in tandem to enable clustering of OTUs from millions of unique reads. These clustering techniques can be classified into three main categories based on the approaches used, these are reference, de novo, and open reference based clustering. These are described in more detail below and summarised in Table 1.2.

Reference based clustering When clustering sequences to OTUs the ultimate aim is to determine some taxonomic identity for the reads within the OTU. This is often achieved by selecting the most abundant read in an OTU as a representative sequence. This can then be aligned to one of the existing databases of prokaryotic rRNA reads to identify the likely taxonomic lineage that the representative read, and hence OTU, belongs to (Nguyen et al. 2016). This is common practice to assign taxonomy to OTUs in 16S rRNA gene sequencing studies.

The process of OTU clustering can be sped up by aligning the reads directly to the limited set of reads in the reference database rather than carrying out prior clustering, which requires comparisons between all the reads in the data. This approach is called reference or closed reference clustering (Navas-Molina et al. 2013). Clustering/alignment to the reference database can be carried out with complete pairwise alignments or using greedy algorithms (Figure1.6) to further speed up the process. Reference based clustering has the benefit of being fast, computationally less demanding as only the reference reads must be loaded to memory, guarantees taxonomic assignment will be known for the resultant OTUs,

and provides a method to ensure consistency across experiments using the same database. However, it will not enable detection of novel sequences absent from the database so can under estimate diversity (Rideout et al. 2014), and relies on the quality of the reference database used.

De novo clustering Clustering of OTUs without a reference is termed de novo clustering. If the number of unique reads are low, for instance when using a lower throughput sequencing technology or a low diversity microbiome, it is possible to de novo cluster amplicon sequencing data by carrying out all pairwise comparisons and using hierarchical clustering (Schloss and Handelsman 2005). This is generally not possible using Illumina sequencing from the gut microbiome, which will contain a high number of unique reads. As a result almost all approaches to de novo OTU clustering rely on greedy algorithms and several different algorithms have been developed to this end (discussed in Chapter 5) (Edgar 2010; Rognes et al. 2016; Mercier et al. 2013; Mahé et al. 2014). Even using greedy approaches, de novo clustering is more computationally intensive than reference based clustering as it must load all the data in memory and consider a higher potential number of comparisons. The de novo OTUs generated in one experiment are also not directly comparable to other studies. However, de novo clustering is able to capture novel sequence not in reference databases so can produce more accurate estimates of microbiome diversity (Rideout et al. 2014). Although, as with reference clustering, assignment of taxonomy to OTUs still depends on the quality of the reference database used.

Open reference clustering One final approach to OTU clustering aims to reap the computational benefits of closed reference clustering with the novelty captured by de novo approaches. Termed open reference clustering, it was developed by the group who created QIIME, a popular software wrapper for algorithms used in 16S rRNA gene sequencing analyses (Navas-Molina et al. 2013; Rideout et al. 2014). The concept is that reads are assigned to OTUs using closed reference clustering followed by de novo clustering on reads that are not found in the reference database. This performs faster than de novo approaches, but mixes approaches and algorithms so leads to an admixture of OTU types.

Alternative approaches More recently, several studies have also proposed threshold free alternatives to define OTUs. Not only do they negate the use of arbitrary thresholds but can also avoid the clustering of sequences altogether. Two examples include the DADA2 and minimum entropy decomposition algorithms (MED). DADA2 models the error profile of sequencing runs from sequencing quality scores, it then uses this model to correct sequencing errors. This enables the use of dereplicated reads without clustering (Callahan et al. 2016). MED, an extension of a previous approach called oligotyping, uses the entropy

Table 1.2: A comparison of OTU clustering approaches.

Clustering Approach	**Requires Reference?**	**Pros**	**Cons**
Reference based	Yes	Lower computational requirements, Fast, Comparable across datasets, All OTUs have some known taxonomy	Misses novel diversity (not in reference)
De Novo	No	No prior assumptions, OTUs include all observed diversity	Hard to compare across datasets, Higher computational requirements, Slow, Novel OTUs can have unknown taxonomy
Open Reference	Yes	Faster than de novo, Lower computational requirements, Captures novel diversity	Slower than reference based, Produces mixture of reference and de novo OTUs, Hard to compare across datasets, Novel OTUs can have unknown taxonomy

(or variation) of nucleotides at each position in the variable region queried to identify the most biologically informative differences. This involves aligning reads to identify the positions that contribute the most variance across all the sequences. The assumption is that relevant phylogenetic differences, which are subject to evolutionary pressures, are more likely to be found in the same position than sequencing errors, which will be more randomly distributed along the sequences. Reads can then be sequentially split into groups based on the nucleotides at the most variable positions (Eren et al. 2015).

There have been numerous studies aiming to compare and benchmark the various approaches to OTU clustering (He et al. 2015; Schloss 2016; Westcott and Schloss 2015; Kopylova et al. 2016; Chen et al. 2013). However, there is no gold standard against which to define an optimal OTU. This is discussed further in Chapter 5 where I aim to address this by using heritability estimates in twins, as a measure of how well an OTU represents the underlying biological differences in the microbiota.

Comparison of OTUs across studies

A further limitation to OTU studies arises from their general reliance on heuristic algorithms. The stochastic nature of these approaches can produce variability in OTU definitions between multiple runs on the same data (He et al. 2015). This compounds variability arising from differences in the experimental approaches or reference databases used. These effects can further hinder reproducibility of amplicon-based microbiome association studies (Sze and Schloss 2016). In Chapter 6 I address comparability between disease studies by carrying out multiple association studies for different diseases within the same dataset. In Chapter 7 I overcome analytical differences between datasets by integrating read data from three different cohort studies prior to OTU clustering.

1.3.4 Typical approaches to the analysis of 16S rRNA gene sequencing-based gut microbiome profiles

Once summarised to OTUs 16S rRNA gene-based profiles of the gut microbiome can be analysed in relation to phenotypes of interest. Here, I provide a brief summary of the typical approaches used in such studies. These are further demonstrated in Chapters 3 and 4, where I use established approaches to identify novel gut microbiome associations with host frailty and PPI use.

Diversity measures

Diversity of the gut microbiome is one of the principle features that can be assessed using 16S rRNA gene sequencing. In microbiome research the term diversity principally refers

to either alpha or beta diversity.

Alpha Diversity Alpha diversity is a measure of organism diversity within an individual microbiome. This can be the number of different organisms present, termed richness, or how evenly distributed the abundances of each organism are, termed the evenness. The most simple measure of alpha diversity is simply the number of unique OTUs observed in a sample. There are also more sophisticated diversity indices used in ecology, for instance the Shannon index that considers evenness (Shannon 2001) or the Chao1 index that quantifies richness adjusting for the proportion of rare organisms in the sample (Chao 1984). Alpha diversity indices can be used in association analyses or case-control comparisons to determine if a phenotype of interest influences the diversity of a sample. For instance, several diseases have been associated with a reduced alpha diversity in the gut microbiome (Gevers et al. 2014; Qin et al. 2012; Goodrich et al. 2014b).

Beta Diversity In microbiome research beta diversity refers to the overlap or similarity of two microbiomes in terms of their microbial complement. Various metrics can be used to assess the similarity (or dissimilarity) of two microbiomes. Most simply beta diversity can be quantified by counting the number of shared and unique OTUs between them (the Jaccard index) or can use more complex metrics that consider the relative abundances of each organism in each microbiome (Lozupone and Knight 2005). Calculating beta diversity between all the samples in a study generates a distance matrix representing the compositional similarity between samples. These are well suited for analyses such as PERMANOVA, to identify associations between phenotypes and wide-scale microbiota composition, or principle co-ordinates analysis, which reduces the dimensionality of the OTU data enabling visualisation of the clustering of samples by microbiota composition. The latter approach being widely used to assess clustering by host phenotypes. For instance, distinct clustering between gut microbiome samples from diseased and healthy individuals (Halfvarson et al. 2017) or between microbiota profiles from different body sites (Consortium 2012; Costello et al. 2009).

There are numerous indices and measures for both alpha and beta diversity. However, many do not take into account the taxonomic similarity between the OTUs. Phylogenetic diversity provides a measure of taxonomic alpha diversity (Faith and Baker 2006) and the UniFrac distance was developed to quantify taxonomic similarity between microbiome profiles (Lozupone and Knight 2005). These are summarised in Figure 1.7.

OTU and taxonomic associations

The use of microbiota profiles derived from 16S rRNA gene sequencing is not restricted to high level diversity analyses. It is also possible to probe abundances of specific microbial

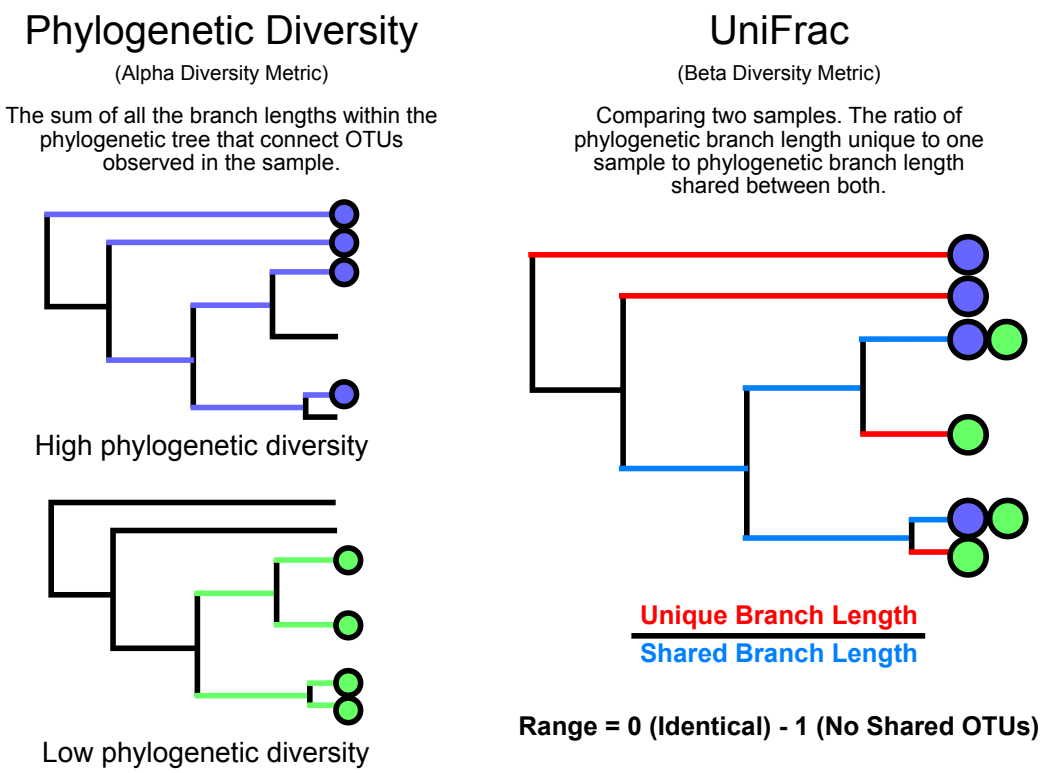


Figure 1.7: Two taxonomy based diversity measures used in microbiome analyses. Shown are the unweighted versions, both can also be adapted to consider the abundance of observed OTUs in samples.

groups in relation to a phenotype of interest. At its finest level these are association analyses with the abundances of individual OTUs. OTU counts can also be collapsed together by shared taxonomic assignment at different taxonomic levels to identify associations with higher level groupings, for instance grouping OTUs from the same families together. There are three main limitations to OTU/taxonomic analyses with 16S rRNA gene amplicon data.

1. OTUs have poor taxonomic accuracy. Taxonomy is assigned to an OTU by alignment of a representative sequence to a reference database. Therefore, the assigned taxonomy may not be representative of all the reads within an OTU (Nguyen et al. 2016). There is also a limit to the accuracy of taxonomic assignment due to both the limits of what has been previously identified (and hence within the reference database) and the resolution that can be delineated from the length of the variable region being probed. As such, OTUs might only be assigned to high taxonomic levels such as phylum. It is also possible to identify multiple OTUs assigned to the same species. This can either occur as a result of the imperfect clustering of greedy approaches leading to multiple OTUs where reads should be grouped together, or as a result of reads that are sufficiently different to cluster separately but similar enough to match reference reads from the same species. Both incomplete taxonomic assignment and duplicate OTU assignments can complicate interpretation of OTU association results: reporting associations with unknown taxa can be uninformative and OTUs with shared taxonomy could have opposing associations with the same phenotype. Both of these limitations were encountered in the analyses in Chapters 3 and 4.

In Chapter 6 I use higher level taxonomic summaries that are less influenced by these effects to generate an index that remains highly correlated with microbiota composition and host health. In Chapter 7, I address the limits of taxonomic accuracy by proposing an alternative analytical approach, collapsing OTUs into interacting communities of OTUs as opposed to using taxonomy based summaries.

2. OTU counts are relative. High-throughput sequencing has different efficiencies across samples in a study. More deeply sequenced samples have a higher chance of detecting rarer OTUs relative to those less deeply sequenced. This can make biologically similar samples appear compositionally different as a result of technical differences. Counts in OTU tables therefore do not represent the number of occurrences in the sample but the number of times that OTU was observed relative to all the other counts observed - OTU counts are compositional data (McMurdie and Holmes 2014). The problem of compositional data effects other fields using high-throughput sequencing based sampling such as mRNA sequencing (Love, Huber, and Anders 2014) and a number of normalisation techniques have been proposed to account for this (McMurdie and Holmes 2014; Weiss et al. 2017). These include converting OTU counts to fold-changes from their geometric mean across all samples in a study (Love, Huber, and Anders 2014; Gloor et al. 2016), or randomly sub-sampling the OTU counts without replacement to a given level to generate even depths across samples, termed rarefying (McMurdie and Holmes 2014). Rarefying of OTU tables has been frequently used within 16S rRNA gene sequencing studies, largely as it is recommended in stages of the popular QIIME pipeline (Navas-Molina et al. 2013). However, by necessity, rarefying of OTU tables results in a loss of information reducing discriminatory power in later analyses (McMurdie and Holmes 2014).

In this thesis I approach OTU normalisation by discarding rarer OTUs that are more influenced by sampling effects, converting raw counts to log transformed relative abundances, and adjusting for the sequencing depth in each sample as a covariate in statistical models. This approach was selected as several of the proposed alternatives have only be compared in differential abundance analyses and not for associations with continuous phenotypes (McMurdie and Holmes 2014; Weiss et al. 2017), it allows simple incorporation of additional continuous covariates, produces easily interpretable results of significance and direction of associations, and has previously been described within the dataset (Goodrich et al. 2014b; Goodrich et al. 2014a). I also use replication of associations in external datasets to validate observations where possible. In Chapter 6 I develop an index to score microbiota composition based on the ratio of different taxonomic counts. As this is based on a ratio of counts it can be applied to raw observations, overcoming issues associated with normalisation between samples.

3. OTU data is high dimensional. Even after collapsing to OTUs there may be thousands of OTUs observed across the samples in a study. This results in a large multiple testing burden when carrying out statistical tests. However, there is also significant inter-correlation between OTUs and parent/child taxonomic levels in a microbiome study that makes many of these tests redundant.

In Chapter 6 I present two approaches to reduce the dimensionality of 16S rRNA gene sequencing data. First, I use correlation patterns between taxa to select a restricted set of representative marker taxa for analyses. Second, I generate a quantification based on these markers that represents wide-scale composition of the gut microbiome in a single index. In Chapter 7 I reduce the number of units for analysis by collapsing OTUs into communities, using the correlations between them to infer interactions.

1.3.5 Alternative approaches to assay the gut microbiome

This thesis is principally concerned with developing approaches to the analysis of microbiota profiles derived from 16S rRNA gene sequencing. As discussed in Section 1.3.3, using such approaches can only provide low resolution taxonomic profiles of a microbiome. Several other culture-independent techniques are available that can provide higher resolution taxonomic information and assay the functional capabilities of the gut microbiome. These include metagenomics, metatranscriptomics, metaproteomics and metabolomics (Tyson et al. 2004; Frias-Lopez et al. 2008; Verberkmoes et al. 2009; Louis, Hold, and Flint 2014). There has also been extension of marker gene approaches to array chips that can provide quantitative information at high taxonomic resolutions by using different marker sequences for each species and/or strain (Paliy et al. 2009). Metagenomic and metabolomic approaches are used in conjunction with 16S rRNA gene sequencing in Chapter 6 and are summarised briefly below.

Metagenomics

Metagenomic approaches to microbiome profiling now typically using shotgun sequencing of microbiome DNA. Complete microbiome DNA is fragmented and sequenced using a high-throughput platform (Sharpton 2014). The resultant short reads can then either be assembled to identify de novo microbial genomes that are then assigned taxonomy (Qin et al. 2010), or reads can be used to estimate taxonomic abundances directly. This is achieved by aligning reads to marker genes specific to individual taxonomic clades (Segata et al. 2012). By assaying taxa at the genomic level metagenomics is able to achieve resolutions as fine as splitting distinct bacterial strains (Scholz et al. 2016). Metagenomics also provides information regarding microbiome function. Genes can be identified in de novo genomes. These and/or the raw reads can be aligned to existing references of known genes. The inferred gene abundances are then typically collapsed to gene homolog groups or functional

pathways, giving the relative abundances of gene functions in the microbiome. It is now also possible to infer function by predicting the presence of functional protein domains from metagenomic DNA sequences (Prakash and Taylor 2012). However, metagenomic approaches can only infer the available functionality from the genes within the microbiome, which may not reflect the active functionality.

It is possible to derive microbiome functionality from 16S rRNA gene sequencing using study databases that have carried out both amplicon-based and metagenomic-based microbiome profiling on the same samples, allowing inference of the correlations between them. The most widely used of these approaches being the PICRUSt package (Langille et al. 2013). Whilst these produce predictions in line with metagenomic assay results, they require use of OTU clustering approaches matching those of the original study, cannot predict functions outside of the training set, and only provide estimates for high level functional pathway abundances (Langille et al. 2013; Aßhauer et al. 2015).

Metabolomics

Metabolomics refers to techniques to assay all the metabolites within an environment. This could be within one microbe, a human tissue, or within samples from a whole microbiome environment, in which case it will contain a mixture of host derived and microbe derived metabolites. Metabolic profiling of a microbiome provides insights into the metabolic functions active within the environment. This in turn provides information regarding microbe-microbe and host-microbe interactions occurring within the community (McHardy et al. 2013). These are important features to study as the gut microbiome has important roles in the metabolism of substrates that cannot be directly utilised by the host, which instead relies on secondary metabolites produced by microbiota (Sonnenburg and Bäckhed 2016). Metabolomic assays require the use of chemical and chromatographic extraction techniques to isolate metabolites in conjunction with either nuclear magnetic resonance or mass spectroscopy to identify them (Lin et al. 2007; Soga et al. 2003).

1.4 Co-occurrence networks in the gut microbiome

1.4.1 Microbial interactions

The composition of the gut microbiome is not only determined by its interactions with the host but also inter-microbial interactions between its members. Microbes can interact via several mechanisms to the benefit of one or both of the parties involved (Figure 1.8) (Faust and Raes 2012). At the microbiome level this results in a complex network of interactions giving rise to functionality that is greater than the sum of its parts (Faust et al. 2012). It has been shown that interspecies interactions can be used to predict the relative

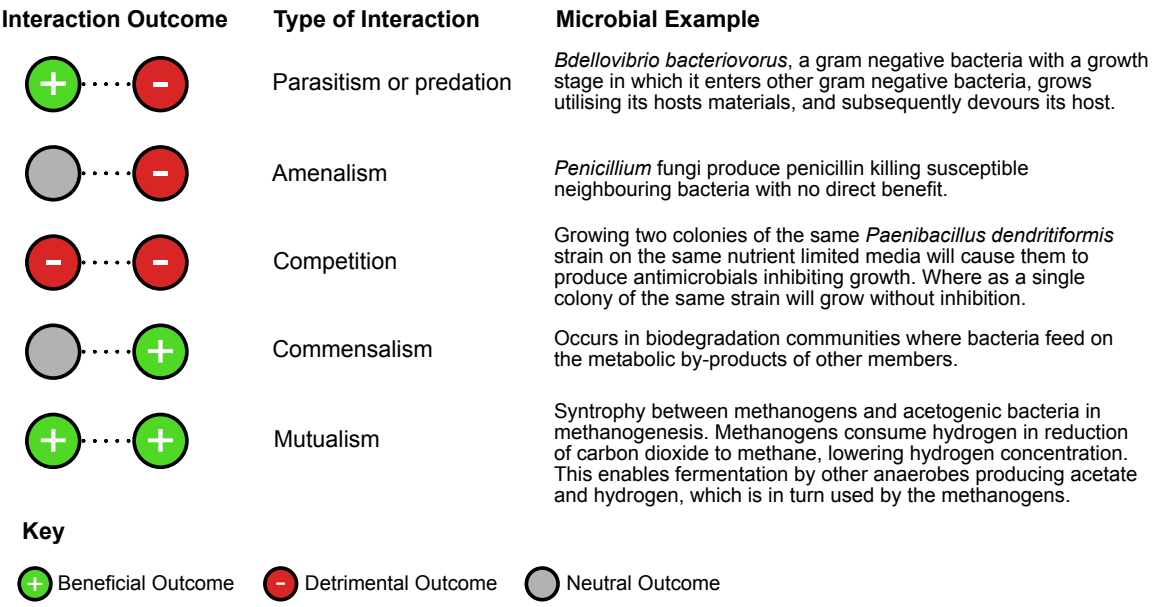


Figure 1.8: Summary of different microbial interaction types with examples. Based on a figure from Faust and Raes 2012. Examples taken from Kadouri et al. 2013; Mougi 2016; Faust and Raes 2012; Morris et al. 2013.

taxonomic composition of mixed microbial communities and can be required to maintain stable community configurations (Friedman, Higgins, and Gore 2017; Larsen, Field, and Gilbert 2012; Konopka, Lindemann, and Fredrickson 2015; Kelsic et al. 2015). For instance, differential antimicrobial production and sensitivity between three species of bacteria can lead to a stable mixed community, whose abundances and spatial conformation can be predicted from their pairwise interactions (Kerr et al. 2002; Kelsic et al. 2015). Whilst this is a simplified model, inter-microbial interactions could similarly influence community structure in the gut. Understanding these interaction networks is important for human health research, as it can provide insights into the mechanisms driving the differences in gut microbiota observed with health related phenotypes.

The gut microbiome contains a diverse range of micro-organisms. Interactions between bacteria can be assayed using pairwise cultures (Friedman, Higgins, and Gore 2017). However this is infeasible for gut microbiota given the wide diversity, and hence high number of comparisons, and that a large proportion of taxa cannot be cultured easily (Browne et al. 2016). Furthermore, in vitro approaches don't represent interactions within the context of the host environment. Host properties such as disease will also influence microbiome interaction networks (Baldassano and Bassett 2016). Thus it is desirable to infer interactions from culture-independent methods such as 16S rRNA gene amplicon-based sequencing.

Interactions can be inferred from 16S rRNA gene sequencing-based microbiota profiles through the co-occurrence, or correlation, of taxa across a range of samples (Faust et al. 2012). For instance, we might expect two taxa with a mutualistic relationship to have a positive correlation between their abundances as they benefit from each others presence. Conversely, a negative correlation between two taxa might represent a competitive inter-

action, where the presence of one reduces the fitness of the other. Co-occurrence is an imperfect assay for interactions, correlations could also arise through taxa responding similarly to external factors, for instance two taxa associating through shared niche specialisation (Faust and Raes 2012; Levy and Borenstein 2013), or through technical artefacts, for instance primer bias to different taxa could induce spurious but consistent associations between their abundances (Walker et al. 2015). These are limitations of using observational data and the experimental approach. However, considering these limitations, assaying co-occurrence patterns pairwise between all the OTUs can still be used for approximate inference of the interactions between them.

1.4.2 Quantification of co-occurrence networks from 16S rRNA gene sequencing profiles of the gut microbiota

The compositionality of OTU count data (see Section 1.3.4) means that common statistical approaches to quantify correlation cannot be used (Faust et al. 2012; Friedman and Alm 2012). As the counts of OTUs sum to one within a sample, there is inherent negative correlation between abundances (Friedman and Alm 2012). For example, in the most simple case of a sample with only two OTUs of even abundance doubling the counts of one will half the relative abundance of the other. This induces correlation into the data as an artefact of sequence sampling rather than from biological sources. Several approaches have been designed to accommodate this problem and estimate co-occurrence between OTUs. Some of these are briefly discussed below and in Chapter 7.

Co-occurrence methods accounting for compositional OTU data

A number of methods have been developed that aim to overcome the problems associated with identifying co-occurrence between OTUs. These can take two approaches, either adapting existing similarity measures to address compositionality, or using novel metrics to quantify co-occurrence that are amenable to compositional data. Existing similarity measures are used by the MENA-RMT, and the CoNet/ReBoot approaches.

MENA Microbial ecology network analysis (MENA) uses classical correlation coefficients, such as Pearson, to determine co-occurrence between OTUs but applies a threshold above which interactions are considered real (Deng et al. 2012). The threshold is determined using a random matrix theory (RMT) approach. A similarity matrix is created from pairwise coefficient values, the threshold is then selected where the eigenvalues of the matrix transition from fitting a Gaussian distribution to a Poisson distribution (Luo et al. 2006).

$$y_{ij} = log\frac{x_i}{x_j} \tag{1.1}$$

Figure 1.9: Equation for the log ratio transform. The log ratio transform provides a measure of proportionality y_{ij} between two measurements in a compositional data set (x_i,x_j) without bias to relative abundances

CoNet Faust *et al.* presented an ensemble based approach to microbial co-occurrence, which has since been released as CoNet (Faust et al. 2012). This quantifies associations by combining the results of multiple measures of similarity (for instance Spearman's and Kendall's correlations) and dissimilarity (for instance Bray-Curtis and Kullback-Leibler distances). Correlations are considered significant if they are observed similarly in all methods, combining the strengths and weaknesses of each measure. CoNet further handles compositional data using a technique called ReBoot to determine the significance of the observed correlations. This overcomes the compositional problem by iteratively permuting OTU counts across samples, followed by re-normalisation to relative abundances. Calculation of correlations after each permutation gives an indication of the level of correlation produced as an artefact of compositionality. Correlations above these null thresholds in multiple co-occurrence measures can be considered true.

SparCC SparCC presents a novel quantification of co-occurrence, designed specifically for compositional OTU data (Friedman and Alm 2012). It utilises the log-ratio transform (Figure 1.9), first proposed by Aitchison (Aitchison 1982) and commonly used in compositional analyses (Gloor et al. 2016). The transformation measures the variance in the ratio between units across a study, and is carried out pairwise between all OTUs. SparCC utilises the ratio variance between two OTUs to infer the coefficient of their association, where zero variance would equal a perfect correlation (the ratio of unit 1 over unit 2 is identical in all samples). SparCC coefficients calculated on permuted tables can then be used to assess the significance of the observed correlations. SparCC has been demonstrated to outperform standard Pearson based analysis when applied to OTU data (Friedman and Alm 2012).

MIC A further novel metric is the maximal information coefficient (MIC). This was developed to be able to identify different of types of associations (linear, exponential etc.) without prioritising detection of any particular type. The MIC approach iteratively bins the two continuous variables under consideration and calculates the mutual information between the bins across the two axes. The MIC coefficient of association is found by the minimum number of bins that produce the maximum mutual information. MIC has previously been applied to microbiome data (Reshef et al. 2011; Weiss et al. 2016).

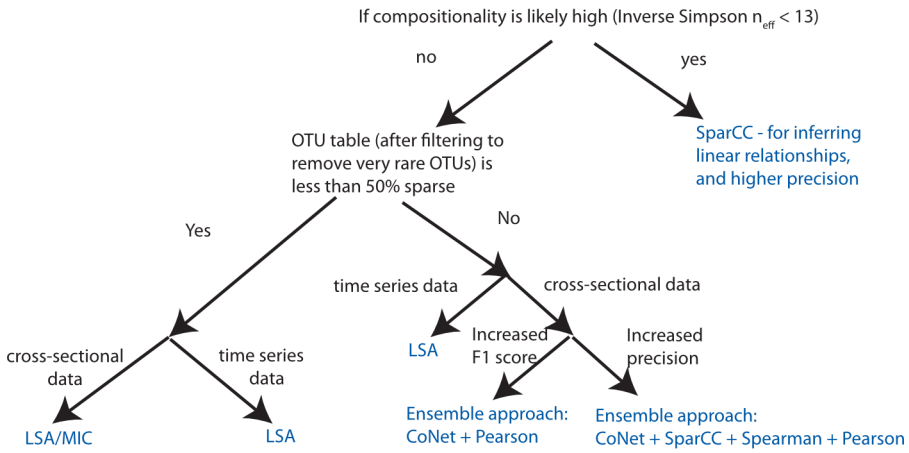


Figure 1.10: Flowchart for selecting the appropriate co-occurrence method to use for microbiome data. From a comparative study by Weiss *et al.* (Weiss et al. 2016).

LSA With suitable datasets, temporal dynamics can also be applied to identify associations between OTUs. In the local similarity analysis (LSA) method defined by Ruan *et al.* covariance of OTU abundances across time series data is used to identify associations between them. LSA can also identify associations where there is a time lag between changes in abundance between OTUs and changes between OTUs and their environment with time (Ruan et al. 2006).

The five approaches described (MENA, CoNet, SparCC, MIC, and LSA) and standard approaches such as Pearson and Spearman correlation were recently compared in their ability to detect OTU associations in a study by Weiss *et al.* (Weiss et al. 2016). The study applied each method to simulated communities based on a range of underlying models and real experimental data. The sensitivity, specificity, and precision, as well as the number of correlations shared across methods, were used to assess the quality of the correlations identified by each method, across a range of different input tables. The final results were summarised in a flowchart advising which method to use based on properties of the starting OTU table and the desired outcomes (Figure 1.10). This was used as a reference to guide co-occurrence detection in Chapter 7.

1.4.3 Defining microbial communities from co-occurrence networks

Interactions inferred from OTU co-occurrence can be considered within interaction networks. Generating networks with the OTUs as nodes and their interactions as edges enables analyses of wide-scale community dynamics. For instance the degree, or connectivity, of OTUs can be used to identify highly connected 'hub' taxa (Faust et al. 2012). Networks can also be compared between groups, for instance between ecosystems (Williams, Howe, and Hofmockel 2014; Faust et al. 2012), populations (Yadav, Ghosh, and Mande 2016), or diseased and healthy states (Baldassano and Bassett 2016; Tong et al. 2013). Furthermore, it is possible to identify sub-networks, or communities, of OTUs within the network.

Within a microbial community we might not expect that all OTUs interact with one another, but rather that they interact within smaller sub-communities, such as those formed in syntrophic metabolism (Morris et al. 2013). In relation to human health research, identifying these communities within the gut microbiome could be particularly useful. There is evidence that observed health-associations with individual gut microbiota can depend on communities of co-occurring taxa (Gevers et al. 2014; Goodrich et al. 2014b; Baldassano and Bassett 2016; Ridaura et al. 2013). Given that collaborative communities are likely a functional unit of the gut microbiome, being able to define them and analyse them in relation to host health could be more relevant that studying whole microbiome or individual taxa level associations. Furthermore, collapsing OTUs to communities would provide a further method to reduce the high dimensionality of gut microbiota data. This is the aim of the method described in Chapter 7, where I present a novel approach to define robust communities within co-occurrence networks of the gut microbiota.

Community detection

Most simply, communities can be identified within networks where there are clearly distinct groups with interactions within but not between them. However, for the most part, large complex data sets, such as microbiota profiles, will not have such clearly defined structures (Faust et al. 2012). Even when all units are within one connected graph it is still possible to identify natural partitions within the structure. There can be more densely connected areas interlinked by more sparse connections. These densely connected areas represent communities (Newman 2006). Identifying/defining these areas within a network is termed community detection. Whilst community detection has only been considered in microbiome networks more recently (Baldassano and Bassett 2016; Tong et al. 2013; Duran-Pinedo et al. 2011), community detection is an established part of graph theory with numerous algorithms having been developed to identify optimal communities, often for applications within social networks (Newman 2004). Below I detail one such approach modularity maximisation, in particular the Louvain algorithm for modularity maximisation.

Modularity Modularity was defined by Newman and Girvan in 2004. When the nodes within a network are assigned to a number of different communities, modularity provides a measure of the quality of the community definitions across the whole network (Figure 1.11). A simple score of the quality of community partitioning is the number of edges found between communities. The best partitions would be expected to minimise the number of edges between communities. Modularity is an extension of this concept, but instead considers if the number of edges within communities is greater than the number of within edges that would be expected by chance. Thus, not only does modularity provide a measure of how well community partitions fit a network structure, but also an indication that there

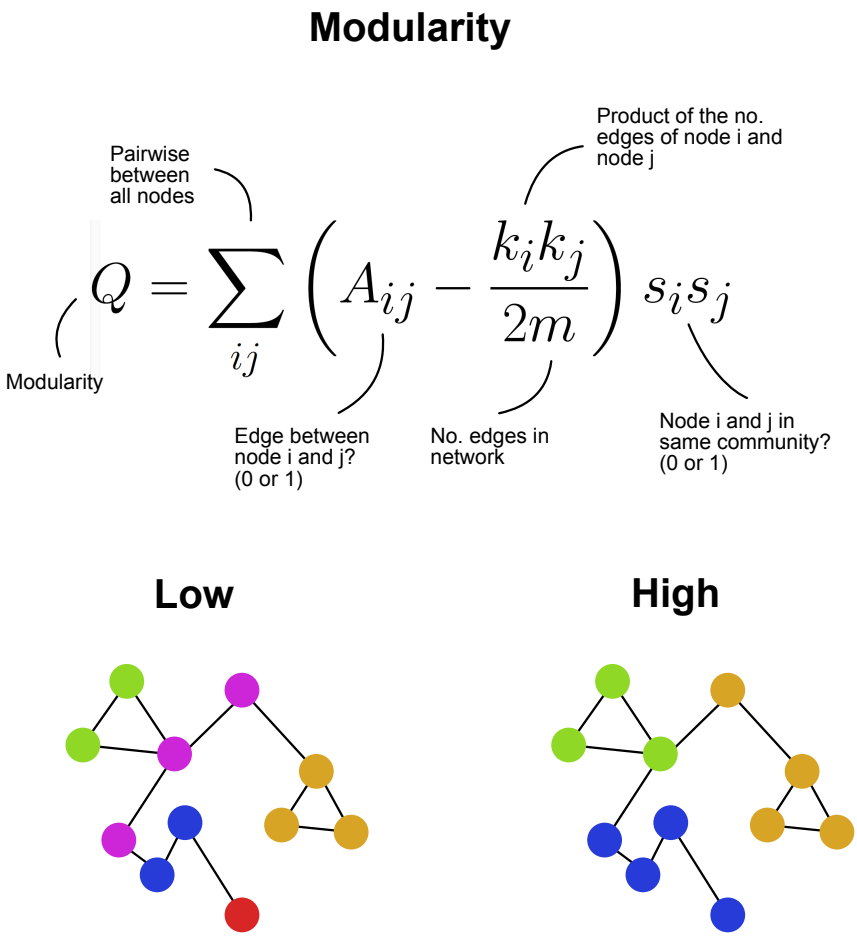


Figure 1.11: The equation to estimate modularity from a partitioned network, with graphical examples of community definitions producing high and low modularity on the same network. (Newman 2006).

is more community structure present than would be expected by chance (Newman 2006).

Louvain Modularity Maximisation Modularity allows scoring of communities that have been detected but not a method to detect and define the communities. However, modularity is often used as the end point of algorithms that define communities by maximising modularity. Several modularity maximisation algorithms have been developed (Newman 2006; Mucha et al. 2010; Blondel et al. 2008; Chen, Kuzmin, and Szymanski 2014). One approached is to iteratively reassign nodes to different communities and only retain changes that produce a positive change in overall modularity. However, even on small networks comparing all possible groupings quickly becomes computationally impossible. A fast and widely used heuristic algorithm for modularity maximisation is the Louvain approach (named after the institute of the developers). This optimises partitioning by first maximising modularity within small local communities around individual nodes. These smaller optimised communities are then considered as nodes and used in a wider analysis, iteratively combining them to form the final global community definitions (Blondel et al. 2008). The Louvain approach is able to identify high quality community definitions within very large networks more quickly than other modularity maximisation approaches (Blondel

et al. 2008).

Community detection in microbiome studies

There have been previous studies that have aimed to quantify communities within microbiota co-occurrence networks. A study by Tong *et al.* used SparCC to generate co-occurrence networks from 16S rRNA gene sequencing from intestinal biopsies of 32 IBD patients and 32 healthy controls (Tong et al. 2013). Within the co-occurrence network they identified 5 genus level communities (or modules) that were conserved across all individuals and 2 that were associated with IBD. To define these modules they used the package WGCNA (weighted gene co-expression network analysis).

WGCNA was developed to identify modules of co-expressed genes in RNA sequencing studies. Unlike the methods previously described, within WGCNA genes (or OTUs) are not assigned to communities uniquely but can have a weighted contribution to each (Langfelder and Horvath 2008). This might be more representative of the true biological nature of complex systems like the gut microbiome, where units contribute to multiple communities in different ways and to a variable extent. However, whilst there are methods to compare these weighted communities across samples, it is simpler to compare communities where units are assigned to single groups. For instance, in the case of gut microbiota studies, OTUs could then be collapsed to communities providing direct community relative abundances for comparisons. This was demonstrated in a recent study using a Louvain modularity maximisation approach to community detection, comparing gut microbiome samples from IBD patients and healthy controls (Baldassano and Bassett 2016). Furthermore, there are instances where it might be desirable to have distinctly defined communities. For instance should a community be associated with a health effect, it would be possible to try and recreate synthetic communities to replicate its effects based on its distinctly defined members.

In summary, modularity provides a measure of the significance of community definitions, it also assigns OTUs to single communities (providing the benefits outlined above), and it can be applied efficiently in large networks. For these reasons, I chose to use Louvain modularity maximisation in the approach outlined in Chapter 7, where I describe a robust method to identify biologically representative communities within 16S rRNA gene sequencing profiles.

1.5 Aims

Aim

The overall aim of this thesis is to develop novel approaches to analyse gut microbiota profiles derived from 16S rRNA gene sequencing to improve their relevance to human health research.

I aim to explore potential solutions to some of the inherent limitations of OTUs discussed previously, including the high dimensionality of OTU datasets, their poor reproducibility, and their limited representation of biologically meaningful units. These are addressed within the following, more specific, aims.

Specific Aims

- In Chapters 3 and 4 I aim to use simple analytic approaches to enhance gut microbiota association studies with the phenotypes frailty and PPI use.
- In Chapter 5 I aim to determine the OTU clustering approach that generates the most biologically relevant OTUs. To this end, I present a comparison of different reference approaches, greedy algorithms, and similarity thresholds using heritability as a novel biologically motivated measure of OTU quality.
- In Chapter 6 I aim to address the poor reproducibility of 16S rRNA gene sequencing studies by carrying out multiple disease association analyses within a single cohort. I also tackle the high-dimensionality of OTU data by presenting an approach to identify a minimal set of marker taxa for analyses, based on the inter-correlations of gut microbiota. Furthermore, I present a novel approach to quantify the gut microbiota using a single index that is representative of both host health and wide-scale taxonomic composition.
- Finally, in Chapter 7 I describe an approach to identify OTU communities within gut microbiota co-occurrence networks. This aims to address the high dimensionality and limited biological relevance of OTU data, by reducing counts to units reflecting biological relationships between OTUs. I demonstrate the robustness and analytical benefits of this approach using comparisons across geographically diverse populations.

Summary

Overall, this body of work describes approaches that address a range of limitations associated with the use of 16S rRNA gene sequencing in gut microbiota studies. Application of these to gut microbiota data from the TwinsUK cohort identifies several novel phenomena demonstrating their utility. These methods should facilitate future analyses utilising 16S rRNA gene sequencing to investigate the gut microbiota in relation to human health.

Chapter 2

Materials and Methods

In the following chapter I provide an overview of the TwinsUK cohort from which the majority of data used in this thesis were derived. I also provide an overview of the laboratory procedures used to generate 16S rRNA gene sequencing data. Finally, I provide a brief overview of some of the methods and techniques that are used throughout. More detailed and study specific data and methods are described within appropriate chapters, several of which are principally concerned with methods development.

2.1 TwinsUK

2.1.1 TwinsUK database

TwinsUK is a population based cohort of volunteer twins managed by the Department of Twin Research and Genetic Epidemiology at King's College London. The recruitment area covers the entirety of the United Kingdom. There are thousands of active individuals registered within the cohort, which includes roughly equal proportions of monozygotic (MZ) and dizygotic (DZ) pairs. The cohort has been active for 25 years and there has been continuous recruitment throughout this period. The majority of the twins are older females. This is due to historical recruitment in the cohort whose original aim was to study osteoarthritis and osteoporosis in women (Moayyeri et al. 2012). A summary of demographics for the participants with 16S rRNA gene sequencing-based gut microbiota profiles is shown in Figure 2.1.

Data is collected from members of the cohort in two ways. Firstly, volunteers are invited to clinical visits at St. Thomas hospital London every 3-4 years. Here they are measured for a range of baseline variables such as weight and height, and provide biological samples, which can include blood, urine, faecal, and skin biopsies. These have been used to assay a range of 'omics' including genotyping of individuals, blood and faecal metabolomics, immunotyping, transcriptomics, and faecal microbiota profiling (see Section 2.2). Volunteers are also

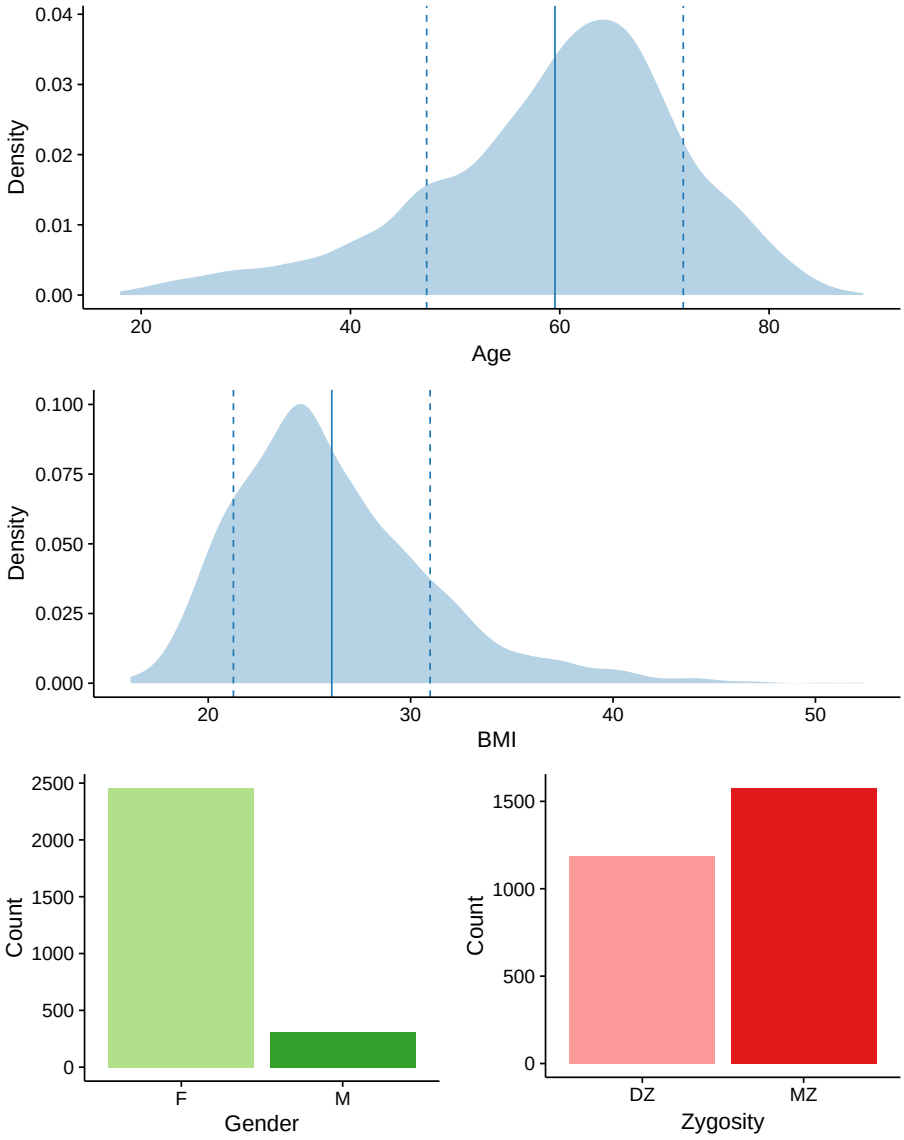


Figure 2.1: Summary properties of the 2764 individuals with 16S rRNA gene sequencing of their gut microbiota within TwinsUK. In the age and BMI density plots the mean and one SD from the mean (dashed) are shown with vertical lines.

subject to a range of physical and mental examinations at clinical visits. Some of these have been assessed consistently throughout the cohort's history and some have only been carried out for specific studies in more limited runs. Secondly, data is obtained via questionnaires. These can be carried out by post, online, or filled out and brought to visits. These assay a range of health and lifestyle factors including ongoing disease status and medication use. There have also been more targeted questionnaires aimed at profiling specific phenotypes, for example food frequency questionnaires used to profile dietary habits.

Throughout this thesis I use a wide range of both visit and questionnaire data. The nature of the cohort means that combining these data often requires sub-setting of larger data sets to account for the incomplete overlap between data collection. There can also be differences in time of data collection between phenotypes.

2.1.2 Heritability estimation

The strength of a twin study is that an individuals co-twin provides a natural control for factors shared within the pair. These include prenatal effects, early life effects, and shared genetic effects. In the case of the latter, MZ twins will share 100% of their genomic complement whilst DZ twins will share on average 50%. Using this knowledge we can observe the differences in a trait between the twins in MZ and DZ pairs and estimate how much of its variance is attributable to additive genetic effects (narrow-sense heritability). For instance, if we observed MZ twins to always be concordant for a disease state and DZ twins to be concordant 50% of the time we can deduce that the disease is 100% attributable to additive genetic effects. This concept can be extended to continuous traits using structural equation models to estimate the variance of a trait attributable to additive genetic variance, the shared (or common) environment within pairs, and the environment unique to individuals (Figure 2.2) (Van Dongen et al. 2012). Heritability estimates are used as a methodological quality measure in Chapter 5 and applied to microbiome measures throughout this thesis.

2.2 16S rRNA gene sequencing from faecal samples

2.2.1 Sample collection and processing

Within TwinsUK, gut microbiota profiles have been generated from 16S rRNA gene sequencing of faecal samples. Participants were given a collection kit to isolate samples at home the day prior to clinical visits. Tubes were then stored in a home refrigerator overnight then brought on ice to visits the next day. Alternatively, samples could be posted on ice from home using a next day service. In both instances, on arrival at the hospital samples were transferred to storage at -80°C. Samples were then shipped to Cornell University on dry ice where DNA extraction, library preparation and sequencing was carried out.

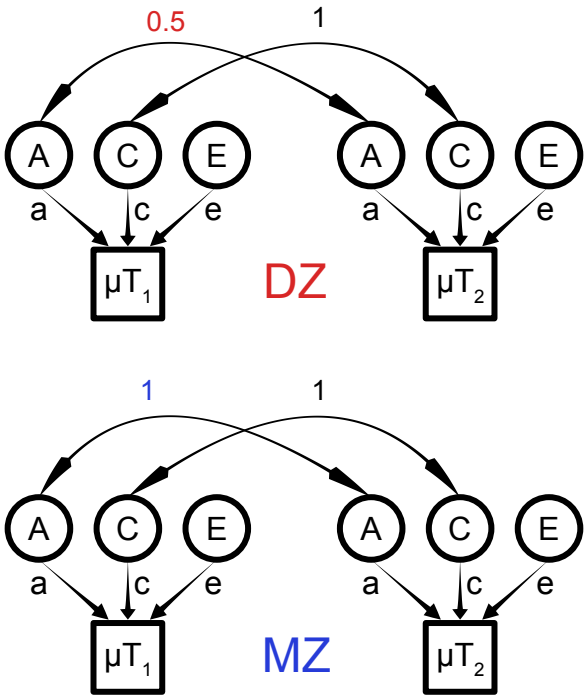


Figure 2.2: Path diagram specifying a structural equation model for a typical ACE model. This provides estimates of the contribution of additive genetic effects (a), common effects within twin pairs (c), and environmental effects (e), to the variance of a continuous trait of interest. Such models can be specified in packages such as OpenMX, which optimise model likelihood to find optimal estimates for the a, c, and e parameters given the observed means and known MZ/DZ relationships (Boker et al. 2011).

DNA was extracted from 100mg of thawed samples using the MoBio PowerSoil DNA extraction kit, which uses both mechanical and physical lysis of cells to release DNA. The V4 region of 16S rRNA genes within each sample were then amplified using PCR, in duplicate, with established barcoded primers unique to the samples (based on the established 515F and 806R primers targeting the V4 region (Caporaso et al. 2011a)). The two PCR products were combined and purified using a magnetic bead approach to isolate long reads and washout contaminants. Individual sample DNA was then diluted to equimolar concentrations and combined prior to sequencing. This was carried out using 250bp paired-end sequencing on the Illumina MiSeq platform (Goodrich et al. 2014b).

2.2.2 Data batches and processing

In total, sequencing of faecal microbiota has been carried out for 2763 individuals with a subset of individuals having multiple longitudinal samples. Collection and processing of these samples was carried out in three main batches over a period of five years from late 2010 to summer 2015. Samples were sent from TwinsUK to Cornell in smaller packages that were aggregated into the three batches. After each batch the existing sequencing data was re-integrated with the previous and re-processed (i.e batch two includes all the sequencing from batch one plus the novel samples). Different processing steps were taken after each batch as the computational approaches within the field advanced. Throughout this thesis

Table 2.1: Batches of 16S rRNA gene sequencing data produced within TwinsUK.

Batch No.	Final Sample Date	No. Twins	Sequencing QC	OTU clustering	Use in Thesis
1	03-2013	951	Paired-ends merged using fastq-join, remove low quality reads (Cornell)	Closed reference clustering with UCLUST and Greengenes 5_13 (Cornell)	Chapter 3
2	07-2014	2025	Paired-ends merged using fastq-join with minumum 200nt overlap, remove low quality reads (Cornell)	Open reference clustering with UCLUST and Greengenes 8_13 (Cornell)	Chapter 4
3	07-2015	2764	As above (Cornell), Chimera filtering with UCHIME (Self)	Various (Chapter 5), De novo clustering with SUMACLUST (Throughout) (Self)	Chapters 5,6, and 7

I utilise data from different batches with different analytical pipelines. In earlier batches pre-processing of 16S rRNA gene sequencing data was largely carried out by researchers at Cornell and in later batches the majority was carried out by myself. A summary of the batches, their size and processing, and how they were used in this thesis is shown in Table 2.1.

2.3 Other microbiome datasets

2.3.1 Metagenomics

Shotgun metagenomic sequencing was also carried out on a subset of faecal samples from 250 individuals. The DNA was extracted as for 16S rRNA gene sequencing at Cornell University. The DNA was shipped on dry-ice to the Beijing Genomics Institute (BGI) in China. Here, DNA samples were fragmented and 100bp paired end shotgun sequencing carried out using the Illumina HiSeq platform. Processing and analysis of these data were carried out at BGI. Reads were filtered to remove those aligning to the human genome. The remaining reads were then assembled to contigs using MOCAT, a metagenomic assembly pipeline (Kultima et al. 2012). Novel genes were identified within contigs using the gene prediction software GeneMark (Lukashin and Borodovsky 1998). Gene abundances were quantified by aligning all reads to these novel genes and an existing in-house non-redundant gene

database created by BGI. For both novel and existing genes functional pathway annotations were made using the KEGG orthology, by aligning translated amino acid sequences to the KEGG functional domain database. Taxonomy was assigned to genes by alignment to existing prokaryotic reference genomes in public repositories. Genes were assigned the taxonomy of an aligned genome when sequence similarity was above a similarity threshold that was different for each taxonomic rank (>65% for phylum, >85% for genus and >95% for species). The KEGG pathway and taxonomic abundances in each sample were found as the summed abundance of their assigned genes (Xie et al. 2016). The KEGG functional abundance table produced from this analysis is used in Chapter 6 to identify functional associations with 16S rRNA gene sequencing derived measures.

2.3.2 Faecal metabolomics

Metabolomic profiling was carried out on faecal samples from 786 members of the TwinsUK cohort. Samples were sent on dry ice to Metabolon inc. who used a proprietary untargeted mass spectroscopy approach to identify and quantify the metabolites present. A subset of 685 these 786 samples also had faecal microbiota profiles. Half were shipped from Cornell, as unused sample from DNA extractions, and the others were sent directly from TwinsUK, where participants had provided duplicate samples at the time of collection. This overlapping dataset is used in Chapter 6 to identify functional associations with microbiota assessed using 16S rRNA gene sequencing.

Chapter 3

Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with host frailty in the gut microbiota

In the following chapter I present an association analysis between host frailty and the gut microbiota, providing a demonstration of typical approaches to 16S rRNA gene sequencing analyses. This includes the use of an existing approach (WGCNA) to identify co-occurrence communities in the gut microbiota. I also demonstrate the use of alpha diversity as a covariate in analyses, enabling the identification of associations independent of wider changes in diversity. These analyses identify several novel gut microbiota associations with frailty. The conclusions of this study, and their context in terms of other chapters within this thesis, are discussed in detail in Chapter 8.

The chapter is presented as a manuscript that was originally published in the journal *Genome Medicine* on the 29th of January 2016. Accompanying supplementary materials can be found in Appendix A. The digital appendix also includes the original PDF file for the manuscript, which has been scaled here to accommodate print margins.

Collaborator Attributions

The frailty index used in this chapter was created by Claire Steves. The replication of significant associations within the ELDERMET cohort data was carried out by Ian Jeffery and Paul O'Toole at the University College Cork. Pre-processing and OTU clustering of 16S rRNA gene sequencing data was carried out by collaborators at Cornell University as described by Goodrich et al. 2014b.

Jackson et al. Genome Medicine (2016) 8:8
DOI 10.1186/s13073-016-0262-7

Genome Medicine

RESEARCH Open Access

Signatures of early frailty in the gut microbiota

Matthew A. Jackson[1], Ian B. Jeffery[2], Michelle Beaumont[1], Jordana T. Bell[1], Andrew G. Clark[3], Ruth E. Ley[3], Paul W. O'Toole[2], Tim D. Spector[1] and Claire J. Steves[1*]

Abstract

Background: Frailty is arguably the biggest problem associated with population ageing, and associates with gut microbiome composition in elderly and care-dependent individuals. Here we characterize frailty associations with the gut microbiota in a younger community dwelling population, to identify targets for intervention to encourage healthy ageing.

Method: We analysed 16S rRNA gene sequence data derived from faecal samples obtained from 728 female twins. Frailty was quantified using a frailty index (FI). Mixed effects models were used to identify associations with diversity, operational taxonomic units (OTUs) and taxa. OTU associations were replicated in the Eldermet cohort. Phenotypes were correlated with modules of OTUs collapsed by co-occurrence.

Results: Frailty negatively associated with alpha diversity of the gut microbiota. Models considering a number of covariates identified 637 OTUs associated with FI. Twenty-two OTU associations were significant independent of alpha diversity. Species more abundant with frailty included *Eubacterium dolichum* and *Eggerthella lenta*. A *Faecalibacterium prausnitzii* OTU was less abundant in frailer individuals, and retained significance in discordant twin analysis. Sixty OTU associations were replicated in the Eldermet cohort. OTU co-occurrence modules had mutually exclusive associations between frailty and alpha diversity.

Conclusions: There was a striking negative association between frailty and gut microbiota diversity, underpinned by specific taxonomic associations. Whether these relationships are causal or consequential is unknown. Nevertheless, they represent targets for diagnostic surveillance, or for intervention studies to improve vitality in ageing.

Background

The ultimate goal of ageing research should be to increase health-span, mitigating a lifespan burdened by morbidity. To this end, frailty is a useful indicator of overall health deficit, describing a physiological loss of reserve capacity and reduced resistance to stressors [1, 2]. It predicts adverse health states such as hospitalisation, dependency and mortality better than chronological age [3], and is increasingly important given the ageing global population with UN estimates of 1.2 billion people aged over 60 years by 2025 [4].

The gut microbiome is the collective coding capacity of the >100 trillion bacteria which significantly enriches the metabolism of the human 'superorganism' [5]. It is highly variable between individuals [6], with substantial heritable elements [7], and relatively stable within a healthy adult over time [8]. Inflammation of the gut is associated with disruption of the gut microbiome [9, 10]. Since the gut is the largest interface with external microbes, and frailty is associated with chronic inflammation [11], it is likely that the gut microbiome has a role in frailty.

Age and frailty influence both the composition and function of the gut microbiome in mice [12]. Similarly in humans, significant differences have been observed between the composition of the adult and elderly adult microbiota [13]. When investigating the effects of frailty, significant differences in the abundances of 17 gut microbes were found between 10 highly frail and 13 'low frail' individuals aged over 70 years, from the same care home who shared the same diet [14]. In the larger

* Correspondence: claire.j.steves@kcl.ac.uk
[1]Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Hospital Campus, 3rd & 4th Floor South Wing Block D, Westminster Bridge Road, London SE1 7EH, UK
Full list of author information is available at the end of the article

Eldermet study, the faecal microbiota composition and diversity of 178 older adults varied with level of health dependency. Patients in long-stay continuing care had a less diverse microbiota than short stay, or community dwelling older adults. Dietary intake differed significantly by residence location, and appeared to drive microbiota associations [15]. Across the whole cohort, in various residence locations, the main axis of microbiota composition change correlated with frailty. We recently showed that discrete configurations of microbial taxa can be robustly defined, several of which have distinctive associations with long-term care, frailty and inflammation [16]. These analyses have focused on generally older, more dependent, participants.

Here we aimed to identify associations between frailty and the gut microbiota within a large cohort of younger community dwelling female twins, adjusting for possible confounding effects such as diet, and genetic and environmental factors shared by twins. We describe significant associations between frailty and microbiota diversity and composition, which represent potential diagnostic markers or therapeutic targets for interventions to reduce frailty in ageing.

Methods

Frailty index

Frailty was quantified through the Rockwood Frailty Index (FI), which translates meaningfully from an epidemiological perspective to clinical studies [17]. The FI was created as a proportion of deficits [18], using data from the Healthy Ageing Twin Study [19]. Thirty-nine domains of binary health deficit were created from questionnaire data and clinical tests covering a range of aspects of physiological and mental health (Additional file 1: Table S1).

Microbiota composition

Faecal samples were collected, bacterial DNA extracted, amplified, sequenced and processed as part of a previous study (see methods therein) [7]. Quality filtering and phylogenetic analysis was performed using QIIME 1.7.0 [20]. OTUs were assigned using closed reference clustering with Greengenes v13_5 at 97 % sequence similarity using UCLUST, resulting in the exclusion of 6.2 % of the total sequences that did not cluster to the reference [7].

OTUs that were observed in fewer than 25 % of individuals were not considered for further study. From a total set of 9,840 OTUs (after removing singletons) 16 % passed this threshold, reflective of the sparseness of the data, resulting in a final set of 1,587 OTUs that were used in association analyses. This threshold was applied to focus on associations within abundant OTUs where data sparsity would be less influential on analyses.

Mixed-effects models

The lme4 package in R was used to generate linear mixed-effects models [21]. The FI was root normalised. Technical covariates included sequencing run and number of sequences in each sample. Biological covariates included relatedness (measured by twin family and zygosity), habitual diet (quantified as the first five PCs from food frequency questionnaires (FFQs) previously assigned to different dietary niches within TwinsUK) [22], alcohol intake, smoking status, age and BMI. Sequencing run and familial traits were modelled as random effects. The Anova function in R was used to compare the ability of models with and without the variable of interest to predict the appropriate response.

Alpha diversity was quantified as observed OTU counts and Shannon and Simpson diversity indices. These were used as response variables to assess associations with alpha diversity.

Models were compared with and without frailty for their ability to predict the log transformed relative abundances of OTUs. Zero counts were handled by addition of an arbitrary value (10^{-6}). Modelling was repeated adjusting for alpha diversity to identify associations independent of overall decreases in diversity. The Shannon metric was chosen, as it considers OTU abundance, was not greatly influenced by sequencing depth and was normally distributed. *P* values were adjusted for multiple testing of OTUs, by false discovery rate (FDR) correction using the 'qvalue' package in R [23]. This was similarly carried out on the relative abundances of collapsed species and genera.

Non-parametric correlation

Non-parametric analyses of FI and OTU abundance associations were carried out using the Kendall's rank sum correlation method in R. The root transformed FI was correlated against the residuals of OTU abundance from previously described mixed models excluding FI. The 'qvalue' package was used for FDR correction with significance at 5 %.

Discordant twin analysis

This was used to demonstrate a lack of bias and potential non-genetic effects. FI discordance between twins was determined where the difference in the pair's root normalised FI was greater than one standard deviation of the population's. Paired Wilcoxon signed-rank tests were used to compare the abundance of OTUs between discordant twins. *P* values were Bonferonni adjusted with a significance threshold of $P < 0.05$. This was carried out for MZ twins only, DZ twins only and both combined.

Co-occurrence modules

The R package WGCNA was used to cluster OTUs by co-occurrence [24]. This has been used previously to

select OTU modules from pre-calculated co-occurrence matrices [25]. However, in this instance, we also used WGCNA functions to quantify co-occurrence using Pearson correlation between the log transformed relative abundances of OTUs. The pickSoftThreshold function was then used to identify an adjacency threshold (17) above which the edges of the network created a scale-free topology. This adjacency matrix was converted to a signed topological overlap matrix (TOM) using the TOMsimilarity function. Hierarchical clustering of OTUs was performed from a dissimilarity matrix derived from the TOM to generate a dendrogram. Modules were selected using the cuttreeDynamic function to select 24 groups of co-occurring OTUs, containing at least 10 members each. Eigenvectors representing each module were generated using the moduleEigengenes function, to obtain the first PC from the module's OTU abundance matrix. A dendrogram of module dissimilarity was generated from a Pearson correlation matrix between the eigenvectors, and modules found to be at least 80 % similar were merged. This resulted in 22 final modules, for which eigenvectors were recalculated. The WGCNA cor function was used to calculate Pearson correlation coefficients between these eigenvectors and phenotypic traits.

Eldermet replication

OTU-level metadata associations were investigated in data from the Eldermet study [15], considering 280 elderly individuals with a mean age of 78 years (age range, 64–102 years). The dataset included community-dwelling individuals, outpatient day-hospital visitors, and short-term and long-term care dwelling individuals.

The Eldermet reference sequences were mapped to the reference sequences from all OTUs, both significant and non-significant, that were assigned to species significantly associated with frailty in TwinsUK. Where there was more than one match between the datasets, the highest scoring BLAST result was retained. Only results above 97 % similarity across the intersect of the sequences were used in subsequent analyses.

Mapped OTUs in the Eldermet dataset were tested for significant association to frailty using a DESeq2 statistical model [26], whereby OTUs were said to be significantly differentially abundant between individuals with a Barthel score <10 compared to a Barthel score of 10 or greater in the Eldermet study if their adjusted *P* value was less than 0.05.

Results

Frailty, microbiome and complete covariate data were available for 728 women from the TwinsUK cohort. The mean age of the cohort was 63 years (age range, 42–86 years) [7]. The FI followed the expected gamma distribution (Additional file 2: Figure S2) [18], had a mean of 0.116, with 103 pre-frail individuals (FI >0.20) [27]. Thus, on average, members of this cohort have low frailty indices reflective of their age and community dwelling status.

Frailty negatively associates with gut microbiota diversity

Linear mixed-effects modelling of alpha diversity versus FI adjusting for age, diet, alcohol intake, smoking, BMI and technical covariates, showed frailty had a significant negative association with alpha diversity (Table 1). It was the most influential determinant of alpha diversity for all metrics considered except observed OTU count, which was influenced by the number of sequences (more deeply sampled sequences having more low abundance OTUs).

OTU abundances that associate with frailty

Models were extended to investigate frailty effects on OTU abundances. Significant associations were found with 637 OTUs using FDR 5 % (Additional file 3: Table S3). Negative associations were enriched for Clostridiales, in particular Ruminococcaceae, with *Faecalibacterium prausnitzii* the most abundant species level assignment (21 OTUs, all negatively associated). Positively associated OTUs that could be assigned at the species level included *Eubacterium dolichum* and *Eggerthella lenta*.

Modelling was repeated also including Shannon diversity as a covariate. This was carried out to identify frailty specific associations that did not result from the overall decrease in alpha diversity; for instance, cases where a microbe's abundance is reduced due to reduced abundance of cooperative taxa rather than frailty effects directly. After adjustment for alpha diversity, 22 OTUs remained significant at FDR 5 % (Additional file 3: Table S3). Nineteen belonged to the order Clostridiales; 11 of which were assigned to the family Lachnospiraceae and eight to Ruminococcaceae, although the direction of association was not consistent at the family level. Two OTUs were assigned to the order Erysipelotrichales and one to Coriobacteriales. Three OTUs were assigned species-level taxonomy corresponding to *F. prausnitzii* (FDR q = 0.027 Beta = −0.14), *E. dolichum* (FDR q = 0.013 Beta = 0.17) and *E. lenta* (FDR q = 0.048 Beta = 0.13).

The FI was constructed from measurements taken 0–3 years before faecal sampling. We would expect a monotonic increase in frailty with age [18], so differences in the time between FI quantification and sample collection between individuals should not significantly influence results and only serve to reduce associations. However, to ensure time passage had no effect, OTU level modelling was repeated including the time difference between the

Table 1 Alpha diversity associations with covariates and the frailty index

Variable	Diversity metric					
	Shannon		Simpson		Observed OTUs (n)	
	P value	β coefficient	*P* value	β coefficient	*P* value	β coefficient
Age	2.08E-01	5.19E-02	3.39E-01	3.92E-02	4.68E-01	2.52E-02
Alcohol intake	9.53E-01	2.93E-03	8.74E-01	−8.16E-03	9.87E-01	6.55E-04
BMI	6.66E-01	−1.71E-02	9.70E-01	−1.53E-03	8.98E-01	−4.25E-03
FFQ PC1 (fruit and veg diet)	8.90E-01	5.14E-03	5.25E-01	2.42E-02	9.41E-01	−2.30E-03
FFQ PC2 (high alcohol diet)	7.46E-01	1.54E-02	6.65E-01	−2.11E-02	5.16E-01	2.55E-02
FFQ PC3 (traditional English diet)	6.19E-02	−7.04E-02	8.18E-01	−8.94E-03	9.88E-02	−5.19E-02
FFQ PC4 (dieting diet)	6.17E-01	1.94E-02	9.18E-01	−4.12E-03	7.86E-01	8.63E-03
FFQ PC5 (low meat diet)	2.15E-01	−4.97E-02	2.66E-01	4.57E-02	3.31E-02	−7.08E-02
Frailty	9.54E-07	−1.93E-01	8.85E-05	−1.59E-01	2.73E-06	−1.53E-01
Sequences (n)	1.06E-01	7.87E-02	3.01E-01	−4.60E-02	5.48E-53	6.76E-01
Smoking status	6.49E-01	NA	2.77E-01	NA	7.51E-01	NA

Mixed effects models were created with Shannon index, Simpson index or number of observed OTUs as the response. ANOVA was used to compare models with and without each variable. Comparisons where *P* <0.05 are highlighted

faecal and FI measurements as a covariate. All previously identified OTUs retained significance, with the number of significant OTUs increasing slightly to 672.

Differential OTU abundance between discordant twins

To account for genetics and environmental influences within twins, the difference in abundance of the 22 OTUs significant after alpha diversity adjustment was examined between 111 pairs (65 DZ and 46 MZ) discordant for frailty. Three OTUs showed significantly differential abundance (Fig. 1). An OTU assigned to the genus *Dorea* was the only OTU more abundant in frailer twins ($P = 4 \times 10^{-3}$). Two OTUs were less abundant in the frailer individuals, which were the *F.prausnitzii* OTU ($P = 0.03$) and one assigned to the family Lachnospiraceae ($P = 4 \times 10^{-3}$). No OTUs were significantly different considering the discordant MZ subset only and the *Dorea* OTU was the only OTU significant within DZ pairs alone. In cases where an OTU was not detected in one twin within a pair, absence generally followed the expected direction from previously observed associations.

Frailty associates with species and genus abundance

OTU counts were collapsed by shared taxonomic assignment at the species, genus and family level. The abundances of taxonomic groups were modelled with frailty

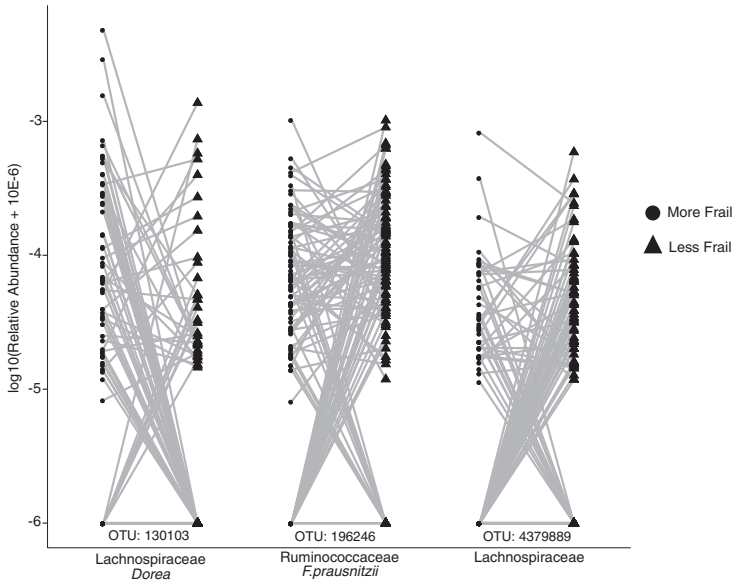


Fig. 1 OTU abundances significantly different between twin pairs discordant for frailty. Pairwise plots of abundance between twin pairs discordant for frailty are shown for OTUs that were significant at $P < 0.05$ after Bonferonni adjustment considering both MZ and DZ pairs. Paired Wilcoxon rank-signed tests were used to compare abundances for the 22 significantly associated OTUs associated with FI after alpha adjustment. Three were significantly different; the *Dorea* OTU was significantly increased in frailer twins ($P < 10^{-4}$), whilst the *F. prausnitzii* and Lachnospiraceae OTUs were significantly lower ($P = 0.001$ and $P < 10^{-3}$, respectively). Note there are overlapping data points where multiple twins had zero counts (log abundance −6) for OTUs

as a predictor (Additional file 4: Table S4). Twelve species traits were significantly associated with frailty at FDR 5 %. The three most significant were *F. prausnitzii*, *E. dolichum* and *E. lenta*. Other collapsed species were largely from the order Clostridiales and negatively associated with FI. *E. dolichum* and *E. lenta* positively associated with frailty. Only *E. dolichum* remained significant after adjustment for alpha diversity.

Twelve genera were FDR significant without alpha adjustment. Ten were negatively associated with FI, seven belonging to the order Clostridiales. The only genera positively associated with FI were *Coprobacillus* and *Eggerthella*, the latter the only genus significant after alpha adjustment. The direction of association with frailty of collapsed species and genera containing *F. prausnitzii*, and *E. lenta* reflected those of their constituent OTUs, whereas the *Eubacterium* genus was not significantly associated with FI (Fig. 2). Two collapsed families were significantly associated with FI at FDR 5 %; both belonged to the class Mollicutes but were non-significant after alpha adjustment.

A number of species were not detectable (classified as absent) within the extremes of frailty, for example, *E. dolichum* and *E. lenta* within less frail individuals. We carried out non-parametric analyses to ensure this effect was not inflating significance in mixed models. This was carried out at the OTU level where absence is most prevalent. Associations between OTU abundance and FI were reassessed using Kendall's rank sum correlations (Additional file 5: Table S5). A total of 138 OTUs were significant at FDR 5 %. All were found in the 637 OTUs identified without alpha adjustment, except seven newly identified OTUs that were all assigned to the family Ruminococcaceae. They also contained 16 of the 22 OTUs significant after alpha adjustment, including the *F. prausnitzii*, *E. dolichum* and *E. lenta* OTUs. This shows that using non-parametric methods, where absence is less influential, the top hits remained significant.

Replication in the Eldermet cohort

We sought to replicate associations using data from the Eldermet cohort [15, 16], a study that used different methods for sequencing and taxonomic assignment and quantified frailty differently using the Barthel index. Eldermet is also a frailer, more elderly and more dependent cohort, including men and women.

To maximise overlap, OTUs were selected within TwinsUK at the species level. That is, all OTUs that

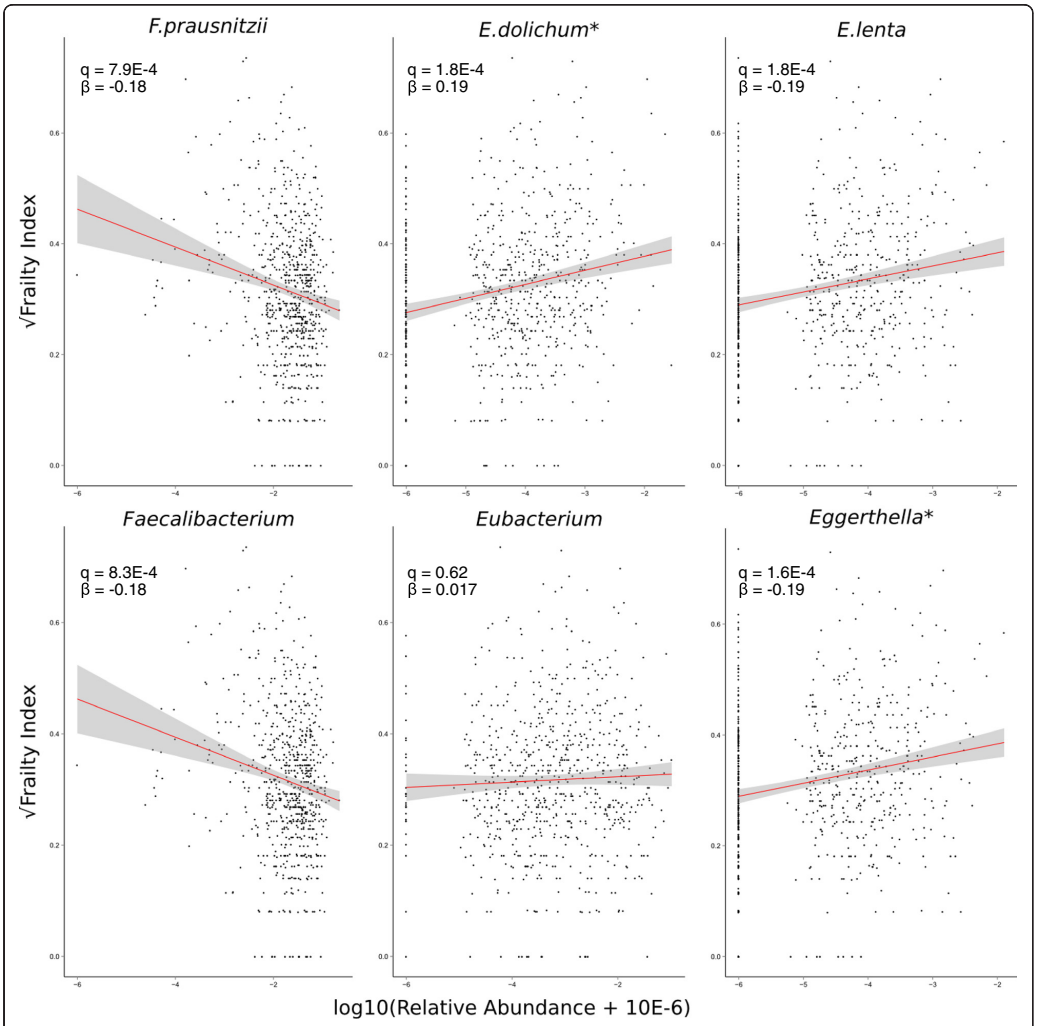


Fig. 2 Species abundance associates with FI. Shown are the unadjusted relative abundances versus FI for all species traits that were significantly associated with FI in mixed effects models, and had complete species name assignment. Below, similar plots for their parent genera. *Represents taxa whose association remained FDR significant after adjustment for alpha diversity. Q-values and coefficients shown are without alpha adjustment. Trends are highlighted by linear regression lines shown in red with 95 % confidence intervals shaded in grey. OTUs contained within these taxonomies retained significance within non-parametric analyses

were assigned to species that were significantly associated with FI within TwinsUK were considered for replication regardless of the OTU's association. Representative sequences from these OTUs within TwinsUK were matched to representatives of OTUs in the Eldermet dataset. Of 638 TwinsUK OTUs considered, 435 mapped to 191 OTUs in the Eldermet cohort. A number of OTUs mapped to the same OTU in the Eldermet set and the remaining 203 had less than the required 97 % similarity to the Eldermet representative sequences and could not be mapped.

Of these 191 mappable OTUs, 96 were significant FI-associated OTUs in TwinsUK. Within the Eldermet cohort, 61 of the 96 were significantly differentially

abundant between individuals with a Barthel score of less than 10 compared to a Barthel score of 10 or greater (adjusted P value <0.05) as defined the by DEseq statistical methodology. The association to frailty was consistent for all but one of these results.

Within the 60 OTUs that replicated across the two studies, 55 were negatively associated with FI and five positively associated with FI. Of the negatively associated, 54 of the 55 were Clostridia/Clostridiales with the final OTU being Mollicutes/RF39.

The five positive associations were *E. dolichum*, *E. lenta*, two Ruminococcaceae and another Clostridia/Clostridiales. This is consistent with observations in the previously published Eldermet studies that showed populations of co-abundant OTUs that contain a large number of unclassified Clostridiales being both positively and negatively associated with frailty and biological ageing [16].

Of the TwinsUK significant variables that failed to reach significance in the Eldermet cohort, 23 of the 35 had an association to frailty that was consistent in direction between the two studies.

TwinsUK OTUs that belonged to significant species but were non-significantly associated at the OTU level, mapped to 95 OTUs in the Eldermet cohort. Of these, 38 had a significant association to frailty in the Eldermet cohort but the association to frailty was inconsistent, with 11 of the 38 showing the opposite association. This shows that the non-significant TwinsUK OTUs have an inconsistent association in the two datasets and so species level associations are poor markers of frailty.

These results highlight the advantage of carefully identifying a set of high quality predictors of frailty using the twin cohort. The significant OTUs returned were consistent in the direction of the association with frailty with the Eldermet observational study.

OTU co-occurrence modules have contrasting associations between frailty and diversity

Bacteria that share functionality or which are interdependent may be taxonomically unrelated but associate similarly with frailty. To overcome this, we collapsed OTUs by co-occurrence, producing 22 co-occurrence modules that were labelled using colour names (the module 'grey' contained unallocated OTUs) (Additional file 6: Table S6). Each module represented groups of OTUs with similar abundance profiles across samples, often with similar taxonomic backgrounds. For example, the module coloured brown in Fig. 3 contained 59 Clostridiales OTUs; these included 14 assigned to *F. prausnitzii* which, upon inspection of the representative sequences, were found to share greater than 97 % sequence similarity, highlighting the limitations of OTU clustering and the ability of co-occurrence collapsing to identify similar features.

Eigenvectors representing each module were used to correlate sample traits with module abundance (Fig. 3). Modules showed the strongest associations with FI and alpha diversity, which contrasted in all cases. Modules significantly negatively associated with FI included those coded magenta, royal blue, salmon, midnight blue and tan (Fig. 3); all consisting of OTUs assigned to Lachnospiraceae, in particular that coded royal blue to the genus *Dorea* and tan to the genus *Roseburia*. Other modules negatively associating with FI were black, blue, light cyan, brown, green, purple and green-yellow. All were assigned to Clostridiales OTUs, particularly Ruminococcaceae with brown and green consisting of largely *F. prausnitzii* OTUs. The only module positively associating with frailty was that coded light yellow, containing 13 Enterobacteriaceae OTUs. Clustering traits by their association with modules separated alpha diversity from the other variables. These formed two clusters of differential association, one containing frailty and the other with correlation patterns more similar to those of diversity. BMI, the low meat FFQ PC and the traditional English FFQ PC all clustered with FI. The alpha diversity like cluster contained the remaining dietary PCs, alcohol consumption and age.

Discussion

Robust associations were identified between frailty and gut microbiota composition. Most notably, FI associated with microbiota diversity. There were also modest associations, both positive and negative, with specific taxa. These findings were robust to adjustment for a range of environmental variables and, in what we believe to be a novel analysis, adjustment for alpha diversity. A number of OTU-metadata associations replicated within the Eldermet cohort.

The distinct negative association between alpha diversity and frailty reflects observations from numerous studies in which predisposition to ill health associates reduced diversity [28–31]. In this study we cannot determine if frailty is the cause, or a consequence of lower microbiota diversity. However, the observed changes in the microbiome could contribute, at least in part, to the detrimental health associated with frailty, particularly in the gut. Should this be the case, maintenance or improvement of gut diversity would be a promising target to encourage health in ageing. This could be achieved through lifestyle interventions to alter factors known to influence diversity, such as diet [32, 33]. This might be particularly effective given the similarity of the observed microbiome associations with diet, BMI and frailty.

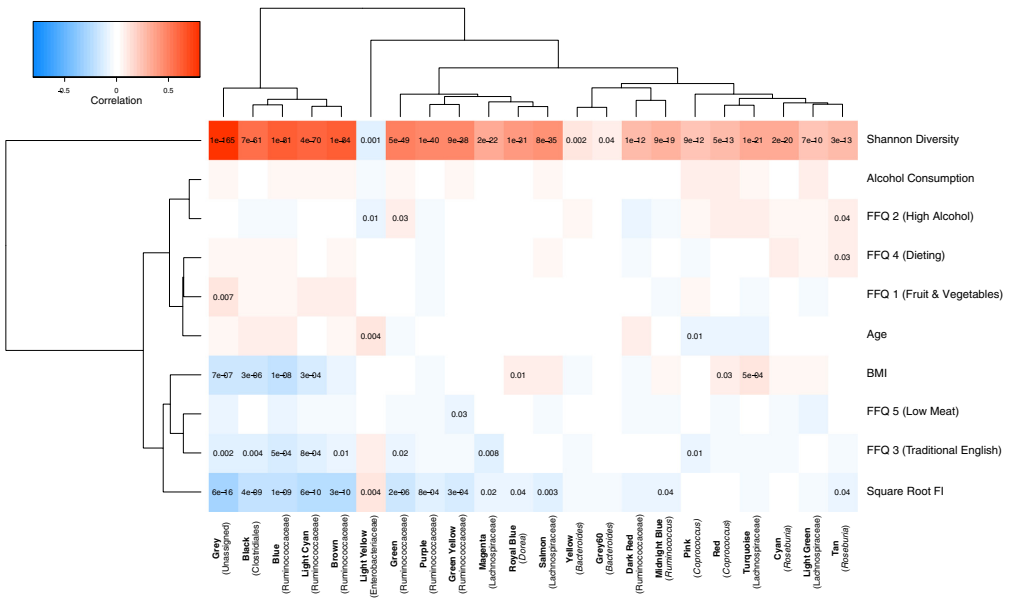


Fig. 3 Correlation of frailty, diversity and model covariates with modules of OTUs collapsed by co-occurrence. OTUs collapsed by co-occurrence into 21 modules labelled by colour (grey containing unassigned OTUs). The heatmap displays the correlation between the module eigenvector (ME, representing the combined module abundance as taken from the first PC of the module's OTUs across all samples) and each variable. Modules and variables are hierarchically clustered using Euclidean distances. Student asymptotic *P* values for significant correlations ($P < 0.05$) are shown. The most common taxonomic assignment within a module is shown next to its colour label

We also observed specific taxonomic associations with frailty in the gut microbiota. These were most apparent at the OTU level where a number of significant associations were also observed within the Eldermet cohort, which utilised differing sequencing and analysis platforms, quantified frailty using an alternate index, and used a frailer population [15]. The robustness of our results is supported by their replication within this contrasting data. However, although they were significant, OTU associations were had modest effect sizes. Further studies will be required to investigate the importance of specific taxa in frailty. Our observations provide motivation and direction for such, and are discussed further below.

A number of *F. prausnitzii* OTUs negatively associated with frailty, reflecting previous observations of reduced *F. prausnitzii* abundance in frail or elderly individuals [14, 34]. *F. prausnitzii* is a key butyrate producer [35, 36], thriving in low pH environments created by other short-chain fatty acid producers [37]. There is evidence that *F. prausnitzii* (and butyrate [38]) can have an anti-inflammatory effect on the gut in mice [39, 40]. The observed negative associations of FI with a sub-set of Lachnospiraceae OTUs also support the role of butyrate producers in suppressing frailty-associated inflammation. Some, but not all, Lachnospiraceae genera are butyrate producers [41]; including *Roseburia*, which we identified as negatively correlated with FI when collapsing OTUs by co-occurrence. These observations warrant further investigation into the effects of gut butyrate production on the progression of frailty.

E. dolichum and *E. lenta* positively associated with frailty in our experiment. The Eggerthella genus contains a number of pathogenic species, including *E. lenta*, that have been associated with gastrointestinal disease [42–44]. *E. lenta* is also known to harbour a cardiac glycoside reductase operon, which can reduce digoxin, to its more inactive reduced lactone dihydrodigoxin [45]. This drug has been commonly used in frail older individuals to control ventricular rate in atrial fibrillation. Our finding should stimulate further research as to whether the efficacy and toxicity of cardiac glycosides may be modulated by *E. lenta* abundance.

Turnbaugh and colleagues sequenced the *E. dolichum* genome after observing that its parent class, Mollicutes,

was associated with obesity in mice [46]. Its genome was enriched for genes involved in simple sugar processing, which was hypothesised to provide an advantage under a Western diet. The increased abundance of *E. dolichum* are likely a consequence of the significant lifestyle changes, particularly dietary, associated with frailty [15]. However, it is possible that changes to microbiome-encoded metabolism are drivers of frailty. Further studies are warranted to distinguish associations with frailty from those with diet and obesity. Longitudinal observational studies would be of particular use to identify the capacity of specific species to predict an individual's trajectory of frailty.

While the observed microbiome associations with frailty were robust to discordant twin analyses and replication in an independent cohort, as a cross-sectional study they are not suggestive of cause. To address this, longitudinal frailty and microbial assessments are planned within the TwinsUK cohort. This study was also limited by the determination of the gut microbiota composition using 16S rRNA amplicon analysis, and would be improved through the use of whole metagenome sequencing which provides more accurate species level assignments and direct functional information [5].

This study focused on abundant OTUs to reduce the influence of data sparsity on the analyses. However, by using a closed OTU clustering approach and discarding less abundant OTUs we have not explored the association of rare and unidentified microbiota with frailty. This may be important in future investigations as there may be novel bacterial units specific to the frail state.

We have also not considered the effects of antibiotics and other medication usage on the microbiome, as there was insufficient data available. These effects could potentially confound the observed associations as frailer individuals are more likely to utilize more, and multiple, medications and factors such as antibiotics are known to influence the gut microbiome [47]. This is also being considered in future studies.

The TwinsUK cohort are younger and more able when compared to previously studied cohorts, so these early associations in theory may not apply to frailer individuals. However, replication in the Eldermet cohort, where antibiotic users were excluded, indicates that in part these findings are not restricted to the less frail individuals within TwinsUK. Use of alternate methods in the quantification of the microbiota and frailty within the replication cohort also suggests that observed associations are independent and robust to the methods used.

Conclusions

We have identified a number of associations between host frailty and the gut microbiota, including modest associations with specific taxonomic abundances and a striking negative association with microbiota diversity. Although more work is required to delineate the direction of effect between frailty and the composition of the gut microbiota, we believe that the associations we have described here provide motivation and direction for such studies. They also provide microbial targets for future investigation, with the ultimate goal to develop the capability to rationally modulate the gut microbiome to improve health in ageing people.

Ethics

Ethical approval for the HATS (Healthy Ageing Twin Study), from which the frailty data were obtained, and microbiota studies within TwinsUK were provided by the NRES Committee London – Westminster. The Cork Clinical Research Ethics Committee provided approval for the Eldermet study.

Data availability

Sequence data for samples within this study are available as part of previously published data under the European Bioinformatics Institute (EBI) accession numbers ERP006339 and ERP006342 [7]. Other data are available for request from TwinsUK.

Additional files

Additional file 1: Table S1 of domains used in the construction of the frailty index. (DOCX 11 kb)

Additional file 2: Figure S2 showing the distribution of the FI in TwinsUK. The distribution of the frailty index in the TwinsUK cohort fits the expected gamma distribution. Histogram of untransformed FI values for the 728 individuals included in the study with a fitted gamma distribution shown in red. (PDF 5 kb)

Additional file 3: Table S3 of significant OTU-FI associations. (XLS 88 kb)

Additional file 4: Table S4 of significant taxonomic associations with FI. (XLS 11 kb)

Additional file 5: Table S5 of OTU associations with FI in non-parametric analyses. (XLS 176 kb)

Additional file 6: Table S6 of OTU module assignments. (XLS 131 kb)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CJS conceived of the study and design, obtained funding for the TwinsUK analyses, and analysed frailty within the population. MAJ devised and carried out statistical analyses of frailty associations with microbiota in TwinsUK, with preliminary analysis provided by MB. Replication of results in data from the Eldermet cohort was carried out and coordinated by IBJ and PWOT. AGC, REL, JTB and TDS were involved in coordination, collection, sequencing, and processing of subjects and samples in the TwinsUK microbiota study; and provided relevant data and advised on analyses in this project. MAJ, IBJ and CJS wrote the manuscript with input from all authors. All authors read and approved the final manuscript.

Acknowledgements
The TwinsUK microbiota project was funded by the National Institutes of Health (NIH) RO1 DK093595, DP2 OD007444. TwinsUK received funding from the Wellcome Trust, European Community's Seventh Framework Programme (FP7/2007-2013), the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. CJS is funded under a grant from the Chronic Disease Research Foundation (CDRF). Science Foundation Ireland (SFI) funds work in IBJ and PWOTs labs.

Author details
[1]Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Hospital Campus, 3rd & 4th Floor South Wing Block D, Westminster Bridge Road, London SE1 7EH, UK. [2]School of Microbiology and Alimentary Pharmabiotic Centre, University College Cork, Cork, Ireland. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA.

Received: 5 October 2015 Accepted: 5 January 2016
Published online: 29 January 2016

References

1. Fried LP, Tangen CM, Walston J, Newman AB, Hirsch C, Gottdiener J, et al. Frailty in older adults: evidence for a phenotype. J Gerontol A Biol Sci Med Sci. 2001;56:M146–57.
2. Howlett SE, Rockwood K. New horizons in frailty: ageing and the deficit-scaling problem. Age Ageing. 2013;42:416–23.
3. Mitnitski AB, Mogilner AJ, Rockwood K. Accumulation of deficits as a proxy measure of aging. ScientificWorldJournal. 2001;1:323–36.
4. United Nations, Department of Economic and Social Affairs, Population Division (2013). World Population Ageing 2013. ST/ESA/SER.A/348.
5. Gill SR, Pop M, DeBoy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, et al. Metagenomic analysis of the human distal Gut microbiome. Science. 2006;312:1355–9.
6. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
7. Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, et al. Human genetics shape the Gut microbiome. Cell. 2014;159:789–99.
8. Rajilić-Stojanović M, Heilig HGHJ, Molenaar D, Kajander K, Surakka A, Smidt H, et al. Development and application of the human intestinal tract chip, a phylogenetic microarray: analysis of universally conserved phylotypes in the abundant microbiota of young and elderly adults. Environ Microbiol. 2009;11:1736–51.
9. Willing BP, Dicksved J, Halfvarson J, Andersson AF, Lucio M, Zheng Z, et al. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. Gastroenterology. 2010;139:1844–54. e1.
10. Levy R, Borenstein E. Metagenomic systems biology and metabolic modeling of the human microbiome: From species composition to community assembly rules. Gut Microbes. 2014;5:37–41.
11. Hubbard RE, Woodhouse KW. Frailty, inflammation and the elderly. Biogerontology. 2010;11:635–41.
12. Langille MG, Meehan CJ, Koenig JE, Dhanani AS, Rose RA, Howlett SE, et al. Microbial shifts in the aging mouse gut. Microbiome. 2014;2:50.
13. Mariat D, Firmesse O, Levenez F, Guimarães V, Sokol H, Doré J, et al. The Firmicutes/Bacteroidetes ratio of the human microbiota changes with age. BMC Microbiol. 2009;9:123.
14. Van Tongeren SP, Slaets JPJ, Harmsen HJM, Welling GW. Fecal microbiota composition and frailty. Appl Environ Microbiol. 2005;71:6438–42.
15. Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature. 2012;488:178–84.
16. Jeffery IB, Lynch DB, O'Toole PW. Composition and temporal stability of the gut microbiota in older persons. ISME J. 2016;10:170–82.
17. Singh I, Gallacher J, Davis K, Johansen A, Eeles E, Hubbard RE. Predictors of adverse outcomes on an acute geriatric rehabilitation ward. Age Ageing. 2012;41:242–6.
18. Searle SD, Mitnitski A, Gahbauer EA, Gill TM, Rockwood K. A standard procedure for creating a frailty index. BMC Geriatr. 2008;8:24.
19. Moayyeri A, Hammond CJ, Valdes AM, Spector TD. Cohort profile: TwinsUK and healthy ageing twin study. Int J Epidemiol. 2013;42:76–85.
20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
21. Bates D, Mächler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015;67:1–48.
22. Teucher B, Skinner J, Skidmore PML, Cassidy A, Fairweather-Tait SJ, Hooper L, et al. Dietary patterns and heritability of food choice in a UK female twin cohort. Twin Res Hum Genet. 2007;10:734–48. doi:10.1375/twin.10.5.734.
23. Dabney A, Storey JD. qvalue: Q-value estimation for false discovery rate control. R package version 2.2.0. 2015.
24. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
25. Tong M, Li X, Wegener Parfrey L, Roth B, Ippoliti A, Wei B, et al. A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. PLoS One. 2013;8:e80702.
26. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. bioRxiv. 2014; doi:10.1101/002832
27. Rockwood K. A global clinical measure of fitness and frailty in elderly people. Can Med Assoc J. 2005;173:489–95.
28. Cotillard A, Kennedy SP, Kong LC, Prifti E, Pons N, Le Chatelier E, et al. Dietary intervention impact on gut microbial gene richness. Nature. 2013; 500:585–8.
29. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, et al. Richness of human gut microbiome correlates with metabolic markers. Nature. 2013;500:541–6.
30. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature. 2009;457:480–4.
31. Manichanh C. Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. Gut. 2006;55:205–11.
32. Carmody RN, Gerber GK, Luevano JM, Gatti DM, Somes L, Svenson KL, Turnbaugh PJ. Diet dominates host genotype in shaping the murine gut microbiota. Cell Host Microbe. 2015;17:72–84.
33. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 2013;505:559–63.
34. Biagi E, Nylund L, Candela M, Ostan R, Bucci L, Pini E, et al. Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians. PLoS One. 2010;5:e10667.
35. Duncan SH. Growth requirements and fermentation products of Fusobacterium prausnitzii, and a proposal to reclassify it as Faecalibacterium prausnitzii gen. nov., comb. nov. Int J Syst Evol Microbiol. 2002;52:2141–6.
36. Benus RFJ, van der Werf TS, Welling GW, Judd PA, Taylor MA, Harmsen HJM, et al. Association between Faecalibacterium prausnitzii and dietary fibre in colonic fermentation in healthy human subjects. Br J Nutr. 2010;104:693–700.
37. Duncan SH, Louis P, Thomson JM, Flint HJ. The role of pH in determining the species composition of the human colonic microbiota. Environ Microbiol. 2009;11:2112–22.
38. Vinolo MAR, Rodrigues HG, Nachbar RT, Curi R. Regulation of Inflammation by Short Chain Fatty Acids. Nutrients. 2011;3:858–76.
39. Sokol H, Pigneur B, Watterlot L, Lakhdari O, Bermudez-Humaran LG, Gratadoux JJ, et al. Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients. Proc Natl Acad Sci. 2008;105:16731–6.
40. Miquel S, Leclerc M, Martin R, Chain F, Lenoir M, Raguideau S, et al. Identification of metabolic signatures linked to anti-inflammatory effects of Faecalibacterium prausnitzii. MBio. 2015;6:e00300–15.
41. Meehan CJ, Beiko RG. A phylogenomic view of ecological specialization in the Lachnospiraceae, a family of digestive tract-associated bacteria. Genome Biol Evol. 2014;6:703–13.
42. Wurdemann D, Tindall BJ, Pukall R, Lunsdorf H, Strompl C, Namuth T, et al. Gordonibacter pamelaeae gen. nov., sp. nov., a new member of the Coriobacteriaceae isolated from a patient with Crohn's disease, and reclassification of Eggerthella hongkongensis Lau et al. 2006 as Paraeggerthella hongkongensis gen. nov., comb. nov. Int J Syst Evol Microbiol. 2009;59:1405–15.
43. Krogius-Kurikka L, Lyra A, Malinen E, Aarnikunnas J, Tuimala J, Paulin L, et al. Microbial community analysis reveals high level phylogenetic alterations in

the overall gastrointestinal microbiota of diarrhoea-predominant irritable bowel syndrome sufferers. BMC Gastroenterol. 2009;9:95.
44. Thota VR, Dacha S, Natarajan A, Nerad J. Eggerthella lenta bacteremia in a Crohn's disease patient after ileocecal resection. Future Microbiol. 2011;6:595–7.
45. Haiser HJ, Seim KL, Balskus EP, Turnbaugh PJ. Mechanistic insight into digoxin inactivation by Eggerthella lenta augments our understanding of its pharmacokinetics. Gut Microbes. 2014;5:233–8.
46. Turnbaugh PJ, Bäckhed F, Fulton L, Gordon JI. Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe. 2008;3:213–23.
47. Jakobsson HE, Jernberg C, Andersson AF, Sjolund-Karlsson M, Jansson JK, Engstrand L. Short-term antibiotic treatment has differing long- term impacts on the human throat and gut microbiome. PLoS One 2010;5.

Chapter 4

Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with proton pump inhibitor use in the gut microbiota

Within this chapter I further demonstrate typical approaches to gut microbiota analyses with the use of a binary trait - proton pump inhibitor (PPI) use. I identify several novel associations between PPI use and the gut microbiota that are independent of GI indications warranting their use. I also demonstrate a novel use of the public HMP data to classify the site-specificity of taxa and probe the mechanisms underlying the observed associations. The conclusions of this study, and their context in terms of other chapters within this thesis, are discussed in detail in Chapter 8.

This chapter is presented in the form of a manuscript that was originally published in the journal *Gut* on the 30th of December 2015. Accompanying supplementary materials can be found in Appendix B. The digital appendix also includes the original PDF file for the manuscript, which has been scaled here to accommodate print margins.

Collaborator Attributions

Pre-processing and OTU clustering of the 16S rRNA gene sequencing data was carried out by collaborators at Cornell University. PPI use and disease state data was parsed from questionnaires by Maria-Emanuela Maxan. Replication of significant associations within data from a clinical intervention study was carried out by Daniel Freedberg.

OPEN ACCESS

ORIGINAL ARTICLE

Proton pump inhibitors alter the composition of the gut microbiota

Matthew A Jackson,[1] Julia K Goodrich,[2,3] Maria-Emanuela Maxan,[4] Daniel E Freedberg,[5] Julian A Abrams,[5] Angela C Poole,[2,3] Jessica L Sutter,[2,3] Daphne Welter,[2,3] Ruth E Ley,[2,3] Jordana T Bell,[1] Tim D Spector,[1] Claire J Steves[1]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2015-310861).

[1]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK
[2]Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA
[3]Department of Microbiology, Cornell University, Ithaca, New York, USA
[4]Clinical Age Research Unit, Kings College Hospital Foundation Trust, London, UK
[5]Division of Digestive and Liver Diseases, Department of Medicine, Columbia University Medical Center, New York, New York, USA

Correspondence to
Dr Claire J Steves, Department of Twin Research & Genetic Epidemiology, King's College London, St Thomas' Hospital Campus, 3rd & 4th Floor South Wing Block D, Westminster Bridge Road, London SE1 7EH, UK; claire.j.steves@kcl.ac.uk

Received 6 October 2015
Revised 9 November 2015
Accepted 25 November 2015
Published Online First 30 December 2015

Open Access Scan to access more free content

CrossMark

To cite: Jackson MA, Goodrich JK, Maxan M-E, *et al*. *Gut* 2016;**65**:749–756.

ABSTRACT

Objective Proton pump inhibitors (PPIs) are drugs used to suppress gastric acid production and treat GI disorders such as peptic ulcers and gastro-oesophageal reflux. They have been considered low risk, have been widely adopted, and are often over-prescribed. Recent studies have identified an increased risk of enteric and other infections with their use. Small studies have identified possible associations between PPI use and GI microbiota, but this has yet to be carried out on a large population-based cohort.

Design We investigated the association between PPI usage and the gut microbiome using 16S ribosomal RNA amplification from faecal samples of 1827 healthy twins, replicating results within unpublished data from an interventional study.

Results We identified a significantly lower abundance in gut commensals and lower microbial diversity in PPI users, with an associated significant increase in the abundance of oral and upper GI tract commensals. In particular, significant increases were observed in Streptococcaceae. These associations were replicated in an independent interventional study and in a paired analysis between 70 monozygotic twin pairs who were discordant for PPI use. We propose that the observed changes result from the removal of the low pH barrier between upper GI tract bacteria and the lower gut.

Conclusions Our findings describe a significant impact of PPIs on the gut microbiome and should caution over-use of PPIs, and warrant further investigation into the mechanisms and their clinical consequences.

Significance of this study

What is already known on this subject?
- Proton pump inhibitors (PPIs) are widely, and often over, used but recently have been associated with a number of side effects, including an increased risk of *Clostridium difficile* infection.
- The increased risk of infection may be mediated by alterations to the gut microbiota, as observed with antibiotics.
- Previous studies have demonstrated associations between PPI use and the gut microbiota, but have been limited in size.

What are the new findings?
- In a large healthy twin cohort, we identify significant associations between the composition of the gut microbiota and PPI use.
- The most striking association is an increase in Lactobacillales, particularly Streptococcaceae, in PPI users.
- The strongest associations replicated in a small interventional dataset indicating causality.
- Finally, we show that bacterial families increasing with PPI use are more likely to be pharyngeal, not gut, commensals.

How might it impact on clinical practice in the foreseeable future?
- The observed alterations to the gut microbiota with PPI use may be responsible for the observed increases in infection risk, and therefore provide targets for research to mitigate these risks.
- The potential consequences of these changes are motivation for caution against unnecessary provision of PPIs.

INTRODUCTION

Proton pump inhibitors (PPIs) are used to increase gastric pH by suppressing acid production. They are pro-drugs, only becoming functional in the acidic environment of the stomach. Here, activated PPIs inhibit hydrogen–potassium pumps (H+/K+ ATPases), transmembrane proteins responsible for releasing hydrochloric acid into the lumen of the stomach. PPIs inhibit H+/K+ ATPases by binding covalently to the transmembrane domain, with return of acid production dependent on the turnover of new H+/K+ ATPases once PPIs have left the system.[1]

PPIs are frequently used to treat GI tract disorders such as bleeding peptic ulcers, erosive esophagitis, and gastroesophageal reflux.[2–4] They are also used prophylactically to prevent stress ulcers and to reduce GI toxicity associated with certain medications, including non-steroidal anti-inflammatory drugs, aspirin, and steroids, sometimes despite a paucity of evidence.[5–8] PPIs are one of the most profitable classes of drugs in the world[9]; however, the high cost to healthcare systems has led to investigations into possible over-utilisation. These show that over 70% of PPI prescriptions may be inappropriate,[10–12] with the majority of over-utilisation

stemming from unnecessary stress ulcer prophylaxis in patients who do not meet the evidence-based criteria, and a lack of re-assessment of PPI use in the community.[12]

The use of PPIs has generally been considered safe, with low reported incidences of serious adverse outcomes.[13–15] However, recently a number of side effects have been identified, including nutritional deficiencies, increased risk of bone fracture, and risks of enteric and other infections[16–19]; notably, increased risks of community acquired pneumonia and *Clostridium difficile* infection where PPIs may carry a high risk equivalent to that of oral antibiotics.[20 21]

The term microbiome refers to the ecology and functionality of the microbial population within an environment. Nearly every site of the human body has a distinct microbiome with bacterial composition determined by environmental and inter-microbial influences.[22 23] Using amplification and sequencing of the variable regions of the 16S ribosomal subunit it is possible to profile the taxonomic composition of the microbiome of a given sample.[23] Application of this technique has shown changes to gut microbiota in a range of conditions, from IBD to obesity and frailty.[24–26] Thus, factors affecting the microbiome have the potential to drive important secondary effects on health. For example, alterations to microbial communities caused by oral antibiotics may underlie their association with increased *C difficile* infection,[27] and the same could be true for PPIs.

Previous small-scale case–control studies indicate that PPI use can influence the microbiome, but have been limited by focusing on younger individuals or patients presenting a GI disorder, with some conflicting results.[28–32]

Here we investigate the association between PPI usage and the gut microbiota in the largest study published to date, using 16S rRNA profiling of faecal samples collected from over 1800 healthy elderly twin volunteers, allowing adjustment for environmental and heritable factors influencing both PPI use and the microbiome. We identified significant alterations to the diversity and composition of the gut microbiota in PPI users, a number of which were replicated in an intervention study. We also identified a potential mechanism by which PPIs could induce such changes.

METHODS

Microbiota composition analysis

One thousand and eighty-one faecal samples from the TwinsUK cohort had been sequenced as part of a previous study; a further ~1000 twin samples were collected and processed under the same protocol producing reads with an average length of 253nt after barcode removal.[33] Previously generated sequencing was combined with new data and quality filtering and ecological analysis performed using QIIME.[34] Sequences were collapsed to operational taxonomic units (OTUs) using open reference clustering with Greengenes v13_8 at 97% sequence similarity. The OTU table was then sub-set to samples from twins with PPI usage data for use in subsequent analyses.

Medication and GI indication data

PPI use was self-reported at multiple time points up to 10 years before and including microbiome assessment. Use was scored as positive if an individual had reported usage at any time, even if there was a more recent negative report. This method was chosen, as PPI use is often intermittent, the longevity of any PPI mediated microbiome effects are unknown, and most misclassifications would be non-users appearing as users, which would act to reduce the strength of any observed associations. Positive PPI use was recorded a median of 3 years before microbiota assessment (IQR 0.2–4.7 years).

GI indications were scored similarly based on self-reported or professionally diagnosed indications for PPI prescription. As for PPI use, multiple time points were available and individuals were considered positive if any indication had ever been reported. Positive GI indications were a median of 1.5 years from faecal sampling (IQR 0–3.8 years).

Self-reported antibiotic use within the previous month was recorded at the time of sample collection for the majority of individuals, with drug details provided where applicable. Binary scores created from these data were corrected to reflect reported treatments, removing individuals where the reported drug was not an antibiotic.

Cohort and covariate data collection

Within TwinsUK 1827 individuals had both PPI data and faecal samples. The average age was 62 years (range 19–88 years) and 90% were female. The gender and age distribution resulted from historical study recruitment within the cohort.[35] Physical measurements such as height and weight were measured at the time of sample collection.

Habitual dietary patterns were represented by the first five principle components (PCs) of food frequency questionnaires (FFQs) collected before sample collection. These have previously been shown to account for the majority of habitual diet variance and correspond to dietary types (given the names of fruit and vegetable rich, high alcohol, traditional English dieting and low meat diets, respectively).[36] Frailty was quantified as a Frailty Index (FI) using the proportion of 39 binary health deficits that each individual displayed (see online supplementary table S1) from the Healthy Ageing Twin Study collected in 2007–2010.[35 37] Covariate distributions were analysed using two-tailed Wilcoxon rank sum tests to compare the distributions of PPI users and non-users, with a significance threshold of $p<0.05$.

α Diversity

The 1827 samples had a mean OTU count of 82 130 (s=40 506, range 10 460–380 500). The OTU table was rarefied to a depth of 10 000 sequences and used to generate Shannon, Chao1 and phylogenetic diversity indices, as well as observed OTU counts. One-tailed Wilcoxon rank sum tests were performed to test for lower diversity in PPI users versus non-users,[32] taking a significance threshold of $p<0.05$ on the full set of 1827 individuals.

Mixed effects models were created using the lme4 package in R,[38] with α diversity metrics as the response variable to assess the ability of PPI status to predict diversity. Technical covariates included sequencing run and depth of sample sequencing. Other covariates included family, twin structure (a variable of unique values the same for monozygotic (MZ) but different for dizygotic (DZ) twins), diet (the first five PCs from FFQs), age, body mass index (BMI), FI (root normalised), and GI indication status. The Anova function was used to compare the ability of models with and without PPI status to predict each α diversity metric, using the subset of 1200 individuals with complete covariate data.

OTU and taxonomic associations

Mixed effects models were again used to identify associations between PPI use and OTU and taxa abundances on 1200 individuals having complete covariate data. OTUs present in <25% of individuals were discarded and the remaining counts converted to log transformed relative abundances (with the addition of 10^{-6} for zero counts). OTU abundances were used as response variables with covariates as above also including the Shannon index, to reduce associations with OTU markers of α

diversity. The ability of models including and not including PPI status as a covariate to predict abundance of each OTU was quantified using the Anova function in R. p Values were FDR (false discovery rate) corrected using the 'qvalue' package with a significance threshold of 5%.[39] OTU counts were collapsed by shared taxonomy at all taxonomic levels. Modelling was carried out for each level individually in the same manner as for OTUs. These analyses were repeated within the subset of individuals who had not used antibiotics.

Interventional study replication

To further assess the possible causal link between exposure to PPIs and the observed taxonomic changes in TwinsUK, we re-analysed data from a previously published crossover study. Methods for this study have been described.[31] In brief, 12 healthy adult volunteers not exposed to antibiotics within the previous 12 months each took 40 mg of omeprazole twice daily for 4–8 weeks, and donated stool samples before and after the PPI course. Bacterial DNA was extracted from all samples and the V4 region of the 16S rRNA gene was amplified using a primer set identical to that used in the TwinsUK study. As for the TwinsUK cohort, the Greengenes database was used for final taxonomic assignments. To best compare data, we assessed the taxonomic changes within the samples from immediately before and immediately after 4 weeks of omeprazole. We analysed taxa that were significantly associated with PPI in TwinsUK and present in the majority of individual specimens (>50%) in the intervention study. We assessed the magnitude and directions of within-individual changes using rank-sum tests (when the distribution of data was not normal) or paired t tests. Taxonomies assigned as 'Other' against the Greengenes reference were not included as they were not comparable between sets.

RESULTS

PPI use in the TwinsUK cohort

Intermittent data on self-reported PPI usage and GI health over a ~10 year time span was available for 1827 individuals, comprising 374 DZ twin pairs, 410 MZ twin pairs and 259 singletons; 90% were female, with an average age of 62 years. Within this set, 892 (49%) had reported some form of GI indication for PPIs, 229 (12%) had been prescribed PPIs at some point (only 24 having used PPIs without any GI indication), and 704 (39%) had reported neither PPI prescription nor GI indication.

PPI use is associated with age, BMI, frailty, and diet

A number of covariates were selected. These included: age, diet as quantified using the first five PCs from FFQs, BMI, and frailty. The association of these with PPI use was assessed in the subset of individuals with complete covariate data (n=1200, 175 having use PPIs) (figure 1). PPI users were significantly older ($p<10^{-6}$), frailer ($p<10^{-15}$), and had higher BMI (p=0.002). They were also found to be significantly lower scoring on FFQ PC2 (p=0.0003), a dietary component related to high alcohol intake.[35]

Significantly lower diversity in the gut microbiome of PPI users

There was significantly lower (p<0.05) diversity in the gut microbiota of PPI users compared to those not using PPIs with all diversity indices (figure 2). There was no significant difference, with any diversity metric, between the individuals with GI indications compared to those without.

The observed negative association between PPI use and α diversity did not withstand adjustment for family and twin structure, BMI, age, frailty and GI indication, in the 1200 individuals with complete data (see online supplementary table S2).

PPI use is associated with specific taxonomic abundances

Modelling of OTU abundances against PPI use identified 22 OTUs with significantly lower abundance in PPI users; all were assigned to the phylum Firmicutes. There were 32 OTUs positively associated with PPI use, 20 from the order Bacteroidales and seven assigned to the *Streptococcus* genus. The strongest association was with a *Bifidobacterium* OTU ($q<10^{-4}$, β=0.45), followed by a *Streptococcus* assigned OTU ($q<10^{-4}$, β=0.44) (see online supplementary table S3).

To identify specific taxonomic relationships, modelling was repeated against OTU abundances collapsed by shared taxonomic assignment at various depths of classification (see online supplementary table S4). A summary of significant associations is shown in figure 3.

Seven collapsed species were negatively associated with PPI use and were all assigned to Erysipelotrichales or Clostridiales (except from one Cyanobacteria). At the level of genera, nine were found to be negatively associated with PPI use, which were largely Firmicutes, with members of the family Erysipelotrichaceae being the most significantly decreased. Five families were negatively associated with PPI use, most strongly, Lachnospiraceae (q=0.004, β=−0.35) and Ruminococcaceae (q<0.0007, β=−0.26).

There were 24 species positively associated with PPI use. These belonged to the phyla Actinobacteria, Bacteroidetes, Firmicutes (particularly Lactobacillaceae and Clostridiales) and Proteobacteria. The largest increases observed with PPI use were the species *Rothia mucilaginosa* ($q<10^{-6}$, β=0.51) and *Streptococcus anginosus* ($q<10^{-6}$, β=0.48). We observed 24 genera that were positively associated with PPI use. The most significantly increased were *Rothia* ($q<10^{-5}$, β=0.45) and *Streptococcus* ($q<10^{-6}$, β=0.47). Ten families were significantly positively associated with PPI use, the most significant being Streptococcaceae ($q<10^{-6}$, β=0.46) and Micrococcaceae ($q<10^{-5}$, β=0.46).

Taxonomic associations with PPI use are independent of antibiotic use

Self-reported oral antibiotic usage data were available for 1 month before faecal sample collection for 1039 of the 1827 individuals. Antibiotic use was significantly associated with PPI use within this set ($\chi^2(1, N=1309)=8.88$, p<0.002), where 16% of PPI users had used antibiotics compared to only 8% of individuals who had not used PPIs. To ensure this enrichment was not influencing the observed associations, modelling analyses were repeated within a subset of 705 individuals that had reported no antibiotic use and had complete covariate data (see online supplementary table S5).

At all levels of analysis, from OTU to phylum, results reflected those of the wider set. The number of significant associations was reduced because of the smaller sample size, but the majority of associations were retained, particularly the strongest positive associations with Streptococcaceae and other Lactobacillales, and the negative associations observed with the class Clostridia. These results show that the observed microbiome associations with PPI use are independent of increased antibiotic utilisation.

Significant associations between discordant twin pairs

The influence of PPIs on the microbiota of 70 MZ twins discordant for PPI use was investigated to control for shared environmental and genetic effects (see online supplementary methods).

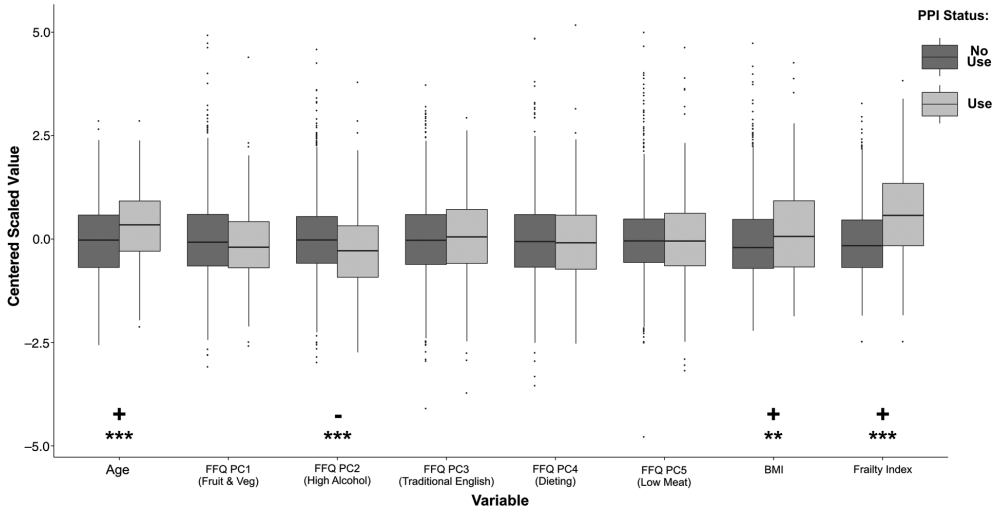


Figure 1 Distributions of covariates included for analysis, compared between proton pump inhibitor (PPI) users and non-users. Wilcoxon rank sum tests were carried out to compare the distribution of covariates in the modelling analysis. All variables were on a different scale so were centred and scaled before plotting. PPI users were older, frailer, had higher body mass index (BMI) and lower scores on the high alcohol food frequency questionnaire (FFQ) principle component (PC). Significant differences are indicated where ***p<0.001 and **p<0.01.

No significant differences in the abundances of any OTU, species or genera were observed between discordant MZs. The Streptococcaceae family had a significantly higher abundance in PPI users within discordant twins (q=0.04, ×2.9 higher), as did its parent order Lactobacillales (q=0.02, ×2.6 higher) (figure 4).

At higher taxonomic levels, significant changes were observed between twins for the classes Actinobacteria (q=0.03, ×1.3 higher in PPI users), Bacilli (q=0.03, ×1.9 higher), and 4C0D-2 (q=0.03, ×0.3 lower), and the phyla Actinobacteria (q=0.04, ×1.1 higher), Cyanobacteria (q=0.04, ×0.33 lower), and Verrucomicrobia (q=0.04, ×0.4 lower).

PPI microbiota associations replicate in an interventional study

From the 96 collapsed taxonomies significant in the TwinsUK set, 63 were found in at least 50% of the interventional set and considered for replication (see online supplementary table S4). Within these, seven were significantly associated with PPI use in the intervention, all increasing in abundance after 4 weeks of PPI use. These were unassigned species belonging to the genera *Streptococcus* and *Granulicatella*, the *Granulicatella* genus, and the families Carnobacteriaceae, Streptococcaceae, Burkholderiaceae, and Corynebacteriaceae. All belonged to the order Lactobacillales except Corynebacteriaceae and Burkholderiaceae, which are from the orders Actinomycetales and Burkholderiales. Within this size-limited intervention the strongest taxa associations, particularly Lactobacilli, appear to be driven by PPI use.

PPI use associated with a higher abundance of pharyngeal bacteria in the gut

Body site of origin of the altered bacteria was investigated to shed light on the mechanism driving observed associations. Data from the human microbiome project (HMP) was used as a reference to determine body site preferences of bacterial families (see online supplementary methods and supplementary table S6). It is worth noting that individual species within each family may be commensals at different and multiple sites. Here we simply aimed to determine overall if families were more frequently identified at particular body sites.

Within the five families found to associate negatively with PPI use in TwinsUK, four were more common in the gut, with two also being found in abundance in the mouth/throat. The exception was Cyanobacteria, which was not biased towards any site in the HMP data. All 10 families positively associated with PPI use (including four replicated in the interventional study) displayed site preference. The six strongest, most significantly associated families were enriched in the mouth/throat with one also in abundance in the skin/nose, one was only enriched at the skin/nose sites, and two were most commonly found at vaginal sites. The only family most common to the gut and with increased abundance with PPI use was Burkholderiaceae. Overall, families with significantly reduced abundance with PPI use were more often found in the gut in the HMP data; while families with significantly higher abundance with PPI use were more often found in the mouth/throat, skin/nose or vaginal sites (likely a result of the large number of Lactobacillaceae commensals found here).[23]

To determine if this trend applied to all families, including those not significantly associated with PPI use, coefficients of association of each family with each site in the HMP data were correlated against families' associations with PPI use in the TwinsUK data. There was a non-significant negative correlation between the association with PPI use and with the gut (ρ=−0.23, p=0.07), and a non-significant positive correlation with vaginal coefficients (ρ=0.2, p=0.12). However, significant positive correlations were observed between the association with PPI use and the association with the mouth/throat (ρ=0.38, p=0.0019) and the skin/nose (ρ=0.36, p=0.003) sites.

DISCUSSION

We have profiled the effects of PPI use on the gut microbiome in by far the largest study to date, and considered a number of

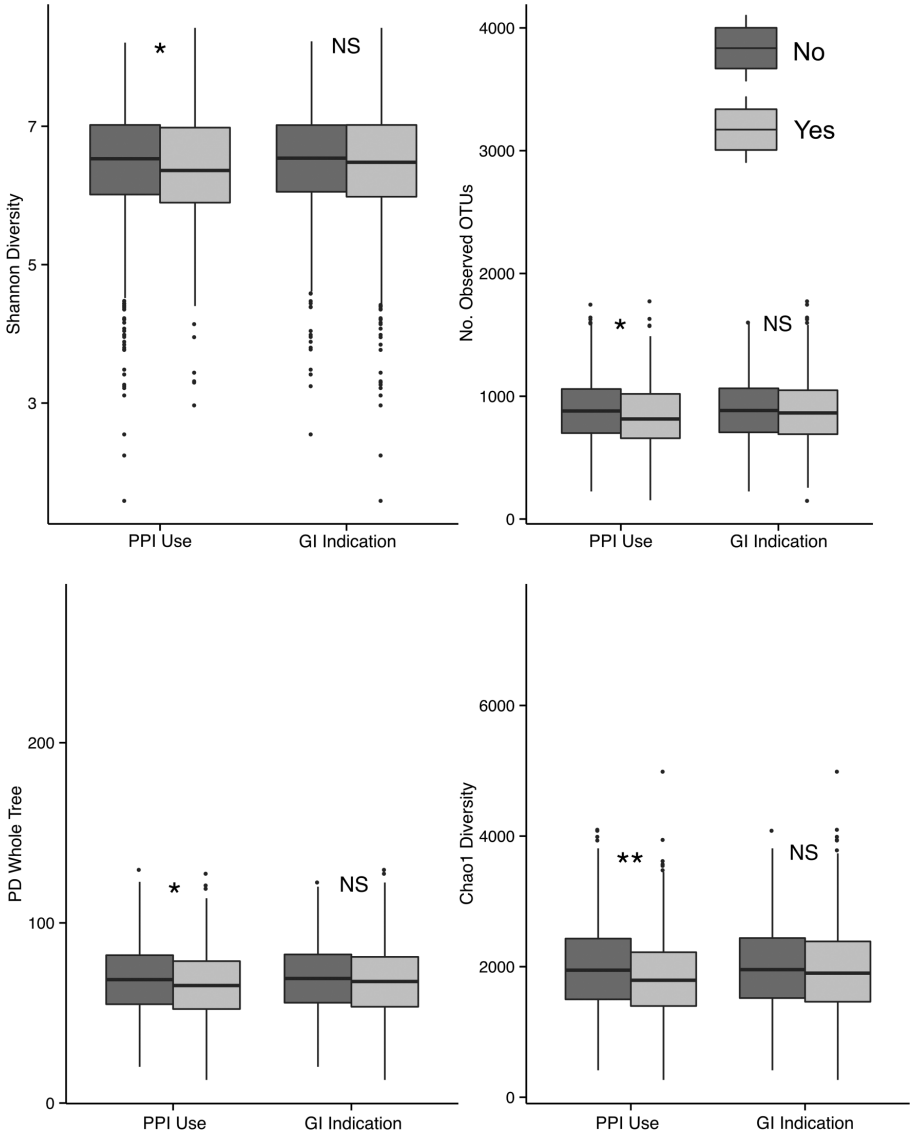


Figure 2 Comparison of α diversity in proton pump inhibitor (PPI) users and non-users, and in individuals with and without GI indications. Four metrics of α diversity were calculated on rarefied samples and one-tailed Wilcoxon rank sum tests carried out to test for significantly lower diversity with PPI use, or with GI indication. Significantly lower diversity was observed for all metrics in PPI users versus non-users (*); no difference was found splitting by GI indication (NS).

possible confounders including host genetics. We have demonstrated that PPI use is associated with an altered composition of the gut microbiota, and a moderately lower diversity. In all three analyses, the large observational study, between discordant twins, and the interventional replication, PPI use was associated with increases in the Lactobacillales order, and in particular the family Streptococcaceae. Further, we show these effects could result from downward movement of upper tract commensals.

We observed a significant reduction in microbial diversity with PPIs. However, it was a small difference and became non-significant after adjusting for covariates. This may be due to confounding effects and/or reduced power of the smaller sample. Variables we found to associate with PPIs, such as BMI, frailty, and antibiotic use, are also known to reduce diversity.[24 26 40] Therefore, it is likely that these factors are partly responsible for the observed lower diversity in PPI users; such confounders were also not accounted for in previous observations of decreased diversity with PPI use.[32] This is further supported by a number of studies where no major changes in diversity have been observed.[28 30]

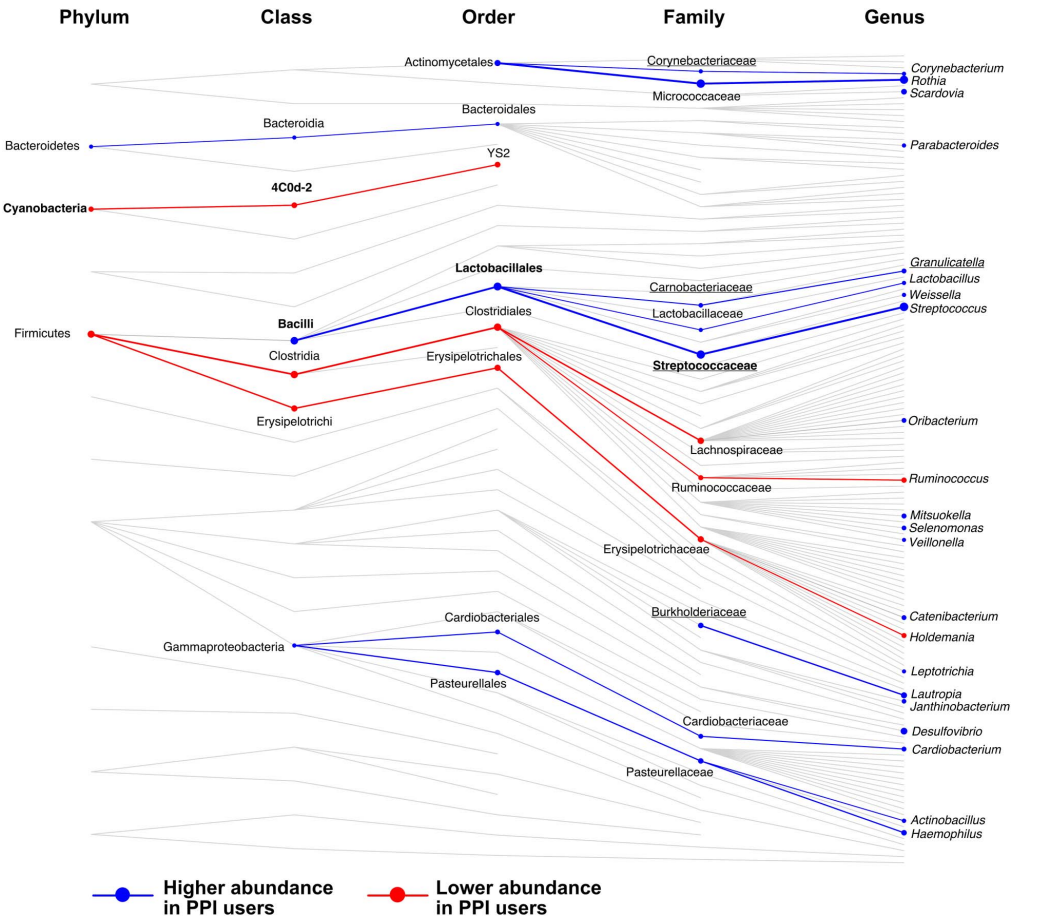


Figure 3 Summary of taxonomic associations with proton pump inhibitor (PPI) use. Shown are all collapsed groups used in taxonomic association analyses that had complete taxonomic assignment (not including collapsed species). Connecting lines highlight the taxonomic relationships between groups (not considering genetic relatedness). Taxa significantly associated with PPI use are highlighted with circles, larger circles representing a larger absolute coefficient of association. Association analyses were carried out at each taxonomic level independently. Taxa at higher abundance with PPI use are shown in blue and at lower abundance in red. Lines connecting taxa of similar association are also coloured and weighted by the average coefficient between the taxa. Names are shown for significant results only. Those in bold retained significance between 70 discordant monozygotic (MZ) twins, and underlined taxa replicated in analysis of interventional study data.

There was a clear association between the composition of the microbiome and PPI utilisation. Collapsing by taxonomic assignment revealed the lineage specificity of these associations, in particular to those containing Streptococcaceae and other Lactobacilli. A number of these associations have been identified in smaller studies. For example, 8 weeks of PPI use was found to increase the abundance of Actinomycetales and Lactobacillales in the oesophagus of 34 patients suffering from heartburn.[29] An increase in *Streptococcus* was also observed with PPI use in a case–control study of 116 children,[30] and previous analyses of the intervention study data utilised within this study also identified similar increases in Streptococcaceae and Micrococcaceae.[31] These studies also identified changes not present in our study, for example, increases in Gemellales, Enterococacceae, and Staphylocacceae.[29 31]

We found families with higher abundance with PPI use to be frequent commensals of the oral, throat, nasal, and skin communities. We hypothesise that under normal circumstances gastric acid acts as a barrier to progression down the GI tract for pharyngeal commensals and environmental bacteria, which are not well adapted to low pH. Treatment with PPIs removes this barrier allowing colonisation by these bacteria further along the GI tract, eventually translating to the detected increased abundance in faecal samples. As observations are based on relative abundances, the observed lower abundance of gut commensals likely reflects the increase in other taxa, rather than a reduction in absolute levels.

PPIs may also act on specific bacterial taxa directly. Previous evidence suggests that they may have antimicrobial action against *Helicobacter pylori*.[41 42] Also, at least one *Streptococcus* species is known to have P-type ATPase transporters belonging

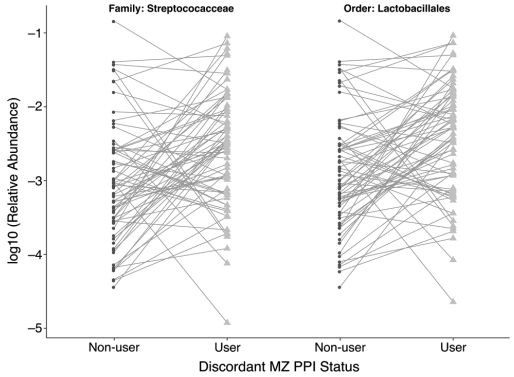


Figure 4 Paired plots of the relative abundances of Streptococcaceae and Lactobacillales within monozygotic (MZ) twins discordant for proton pump inhibitor (PPI) use. These were the only collapsed family and order traits found to be significantly different in discordant MZ twin pairs, both higher in abundance in PPI users.

to the same enzyme family as the human H+/K+ ATPase targeted by PPIs.[43] Bacterial targets for direct PPI interactions could drive species-specific compositional changes.

There are limitations to the study. The TwinsUK data are observational, although our key findings were confirmed between twin pairs and replicated in data from a small prospective controlled trial. PPI use and GI indication were self-reported and over a wide timespan from faecal sampling. However, misclassification of exposures should only serve to reduce the strength of observed associations. We have also omitted duration of PPI use as accurate data were not available, which should be considered in future investigations as it may influence the strength of microbiome associations. However, the lack of duration data would tend to dilute our associations as short and long-term users are classified together. Similarly, we did not consider the effects of withdrawal of PPI treatment, which may be important given that dysbioses resulting from antibiotic use can have long lasting effects.[40]

Antibiotic use was not scored for all individuals within this study; we have also not considered the effects of particular classes of antibiotics, dosage and duration of courses, or antibiotic use before the previous month. However, the robustness of our observations within the subset of individuals who had not used antibiotics shows that they are independent of the effects of recent antibiotic exposure. This study was also limited to faecal sampling. While the observations in the gut are robust, they offer no insight into the distribution of these bacteria along the GI tract. Sampling of multiple sites combined with culture experiments would determine the distribution of living bacteria, and the influence of upper tract community composition on subsequent changes in the gut. In vitro studies will also be required to elucidate whether our observations are driven by pH changes, direct drug interactions, or a combination of both. This will be particularly important to determine if these effects occur with other classes of acid-suppressing medication.

The associations reported here are of clinical importance. *C difficile* affects nearly half a million people in the USA annually,[44] and is known to capitalise on alterations to the normal gut microbiota.[27] The increased risk of enteric infection with PPI use may similarly be mediated through changes to the GI microbiome. It has been shown that a high abundance of *Streptococcus* in the gut predisposes mice to *C difficile* colonisation, while *Lachnospiraceae* are protective.[45] On this basis, we observed taxonomic changes that would be expected to promote *C difficile* infection. Further investigations into the microbiome-mediated determinants of *C difficile* infection will be important to understand how to mitigate the risk associated with PPI use.

A further consequence for consideration is the potential for the GI tract to become a reservoir for potential pathogens at alternate body sites. A significant increased risk of community-acquired pneumonia has been observed with PPIs (relative risk 1.98 for users vs ex-users),[18] and has been observed specifically for *Streptococcus* derived pneumonia.[46] There is speculative evidence of bacterial exchanges between the gastric and lung fluids,[30 47] and depletion of the gut microbiota reduces immune mediated resilience to pneumococcal pneumonia in mice.[48] PPI use has also been shown to increase the risk of spontaneous bacterial peritonitis and overall bacterial infection in patients with cirrhosis and ascites,[19] suggesting PPI use may pose a higher risk to individuals already susceptible to infection and other complications; for example, the elderly and the more frail or more obese individuals, whom our study indicates are more likely to be prescribed PPIs.

The described associations between PPI use and the gut microbiome warrant further research to better understand the driving mechanisms and their consequences, and are a further reason to reduce unnecessary prescribing.

Acknowledgements We thank Dr Frances Williams for her advice in preparation of the manuscript.

Contributors Statistical analyses within TwinsUK were carried out by MAJ. M-EM collated and parsed PPI and GI indication data. MAJ authored the manuscript with the help of M-EM. DEF and JAA carried out replication analysis in the interventional data set. ACP, JLS and DW carried out extraction and sequencing of DNA from samples. JKG carried out pre-processing and OTU picking from sequencing data. CJS conceived the study idea, advised on analysis and coordinated the project. REL, JTB and TDS coordinated the collection and analysis of the microbiota data. All authors contributed to and revised the manuscript.

Funding The TwinsUK microbiota project was funded the National Institutes of Health (NIH) RO1 DK093595, DP2 OD007444. TwinsUK received funding from the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013), the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. CJS is funded under a grant from the Chronic Disease Research Foundation (CDRF). DEF was funded by the National Center for Advancing Translational Sciences (NIH KL2 TR000081).

Competing interests None declared.

Ethics approval Ethical approval for the HATs and microbiota studies within TwinsUK were provided by the NRES Committee London–Westminster.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data are available for request from TwinsUK.

REFERENCES

1 Katelaris PH. Proton pump inhibitors. *Med J Aust* 1998;169:208–11.
2 Bardou M, Toubouti Y, Benhaberou-Brun D, *et al*. Meta-analysis: proton-pump inhibition in high-risk patients with acute peptic ulcer bleeding. *Aliment Pharmacol Ther* 2005;21:677–86.
3 Sharma VK, Leontadis GI, Howden CW. Meta-analysis of randomized controlled trials comparing standard clinical doses of omeprazole and lansoprazole in erosive oesophagitis. *Aliment Pharmacol Ther* 2001;15:227–31.
4 DeVault KR, Castell DO. Updated guidelines for the diagnosis and treatment of gastroesophageal reflux disease. *Am J Gastroenterol* 2005;100:190–200.
5 Yachimski PS, Farrell EA, Hunt DP, *et al*. Proton pump inhibitors for prophylaxis of nosocomial upper gastrointestinal tract bleeding: effect of standardized guidelines on prescribing practice. *Arch Intern Med* 2010;170:779–83.

6 Singh G, Triadafilopoulos G. Appropriate choice of proton pump inhibitor therapy in the prevention and management of NSAID-related gastrointestinal damage. *Int J Clin Pract* 2005;59:1210–17.
7 Jung R, MacLaren R. Proton-pump inhibitors for stress ulcer prophylaxis in critically ill patients. *Ann Pharmacother* 2002;36:1929–37.
8 National Institute for Health and Clinical Excellence. NSAIDs—prescribing issues. 2015. http://cks.nice.org.uk/nsaids-prescribing-issues (accessed 5 Jul 2015).
9 IMS Health Top-Line Market Data. 2014. http://www.imshealth.com/portal/site/imshealth/menuitem.18c196991f79283fddc0ddc01ad8c22a/?vgnextoid=6521e590cb4dc310VgnVCM100000a48d2ca2RCRD&vgnextfmt=default (accessed 6 Jul 2015).
10 Akram F, Huang Y, Lim V, *et al*. Proton pump inhibitors: are we still prescribing them without valid indications? *Australas Med J* 2014;7:465–70.
11 Eid SM, Boueiz A, Paranji S, *et al*. Patterns and predictors of proton pump inhibitor overuse among academic and non-academic hospitalists. *Intern Med* 2010;49:2561–8.
12 Heidelbaugh JJ, Kim AH, Chang R, *et al*. Overutilization of proton-pump inhibitors: what the clinician needs to know. *Therap Adv Gastroenterol* 2012;5:219–32.
13 Fitton A, Wiseman L. Pantoprazole. *Drugs* 1996;51:460–82.
14 Wilde MI, McTavish D. Omeprazole. *Drugs* 1994;48:91–132.
15 Langtry HD, Wilde MI. Lansoprazole. *Drugs* 1997;54:473–500.
16 Valuck RJ, Ruscin JM. A case-control study on adverse effects: H2 blocker or proton pump inhibitor use and risk of vitamin B12 deficiency in older adults. *J Clin Epidemiol* 2004;57:422–8.
17 Leonard J, Marshall JK, Moayyedi P. Systematic review of the risk of enteric infection in patients taking acid suppression. *Am J Gastroenterol* 2007;102:2047–56.
18 Laheij RJF, Sturkenboom MCJM, Hassing, *et al*. Risk of community-acquired pneumonia and use of gastric acid-suppressive drugs. *JAMA* 2004;292:1955–60.
19 Xu HB, Wang HD, Li CH, *et al*. Proton pump inhibitor use and risk of spontaneous bacterial peritonitis in cirrhotic patients: a systematic review and meta-analysis. *Genet Mol Res* 2015;14:7490–501.
20 Janarthanan S, Ditah I, Adler DG, *et al*. *Clostridium difficile*-associated diarrhea and proton pump inhibitor therapy: a meta-analysis. *Am J Gastroenterol* 2012;107:1001–10.
21 Kwok CS, Arthur AK, Anibueze CI, *et al*. Risk of *Clostridium difficile* infection with acid suppressing drugs and antibiotics: meta-analysis. *Am J Gastroenterol* 2012;107:1011–19.
22 Faust K, Sathirapongsasuti JF, Izard J, *et al*. Microbial co-occurrence relationships in the Human Microbiome. *PLoS Comput Biol* 2012;8:e1002606.
23 The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;486:207–14.
24 Turnbaugh PJ, Hamady M, Yatsunenko T, *et al*. A core gut microbiome in obese and lean twins. *Nature* 2009;457:480–4.
25 Willing BP, Dicksved J, Halfvarson J, *et al*. A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes. *Gastroenterology* 2010;139:1844–54.e1.
26 Claesson MJ, Jeffery IB, Conde S, *et al*. Gut microbiota composition correlates with diet and health in the elderly. *Nature* 2012;488:178–84.
27 Owens RC, Donskey CJ, Gaynes RP, *et al*. Antimicrobial-associated risk factors for *Clostridium difficile* infection. *Clin Infect Dis* 2008;46(Suppl 1):S19–31.
28 Tsuda A, Suda W, Morita H, *et al*. Influence of proton-pump inhibitors on the luminal microbiota in the gastrointestinal tract. *Clin Transl Gastroenterol* 2015;6:e89.
29 Amir I, Konikoff FM, Oppenheim M, *et al*. Gastric microbiota is altered in oesophagitis and Barrett's oesophagus and further modified by proton pump inhibitors. *Environ Microbiol* 2014;16:2905–14.
30 Rosen R, Hu L, Amirault J, *et al*. 16S community profiling identifies proton pump inhibitor related differences in gastric, lung, and oropharyngeal microflora. *J Pediatr* 2015;166:917–23.
31 Freedberg DE, Toussaint NC, Chen SP, *et al*. Proton pump inhibitors alter specific taxa in the human gastrointestinal microbiome: a crossover trial. *Gastroenterology* 2015;149:883–885.e9.
32 Seto CT, Jeraldo P, Orenstein R, *et al*. Prolonged use of a proton pump inhibitor reduces microbial diversity: implications for *Clostridium difficile* susceptibility. *Microbiome* 2014;2:42.
33 Goodrich JK, Waters JL, Poole AC, *et al*. Human genetics shape the gut microbiome. *Cell* 2014;159:789–99.
34 Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;7:335–6.
35 Moayyeri A, Hammond CJ, Valdes AM, *et al*. Cohort profile: TwinsUK and healthy ageing twin study. *Int J Epidemiol* 2013;42:76–85.
36 Teucher B, Skinner J, Skidmore PML, *et al*. Dietary patterns and heritability of food choice in a UK female twin cohort. *Twin Res Hum Genet* 2007;10:734–48.
37 Searle SD, Mitnitski A, Gahbauer EA, *et al*. A standard procedure for creating a frailty index. *BMC Geriatr* 2008;8:24.
38 Bates D, Mächler M, Bolker BM, *et al*. Fitting linear mixed-effects models using lme4. *ArXIV e-print; Press J Stat Softw* 2015;67:1–48.
39 Dabney A, Storey JD. qvalue: Q-value estimation for false discovery rate control. R package version 2.2.0 2015 https://www.bioconductor.org/packages/3.3/bioc/html/qvalue.html
40 Jakobsson HE, Jernberg C, Andersson AF, *et al*. Short-term antibiotic treatment has differing long- term impacts on the human throat and gut microbiome. *PLoS ONE* 2010;5:e9836.
41 Hirai M, Asuma T, Ito S, *et al*. A proton pump inhibitor, E3810, has antibacterial activity through binding to *Helicobacter pylori*. *J Gastroenterol* 1995;30:461–4.
42 Nagata K, Satoh H, Iwahi T, *et al*. Potent inhibitory action of the gastric proton pump inhibitor lansoprazole against urease activity of *Helicobacter pylori*: unique action selective for *H. pylori* cells. *Antimicrob Agents Chemother* 1993;37:769–74.
43 Ajdić D, McShan WM, McLaughlin RE, *et al*. Genome sequence of Streptococcus mutans UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci USA* 2002;99:14434–9.
44 Longo DL, Leffler DA, Lamont JT. *Clostridium difficile* infection. *N Engl J Med* 2015;372:1539–48.
45 Schubert AM, Sinani H, Schloss PD. Antibiotic-induced alterations of the murine gut microbiota and subsequent effects on colonization resistance against *Clostridium difficile*. *MBio* 2015;6:e00974.
46 de Jager CPC, Wever PC, Gemen EFA, *et al*. Proton pump inhibitor therapy predisposes to community-acquired Streptococcus pneumoniae pneumonia. *Aliment Pharmacol Ther* 2012;36:941–9.
47 Dumoulin G. Aspiration of gastric bacteria in antacid-treated patients: a frequent cause of postoperative colonisation of the airway. *Lancet* 1982;319:242–5.
48 Schuijt T, Lankelma J, Scicluna BP, *et al*. The gut microbiota plays a protective role in the host defence against pneumococcal pneumonia. *Gut* 2016;65:575–83.

Chapter 5

A comparison of methods to cluster 16S rRNA gene sequences to OTUs using heritability as a measure of quality

In the following chapter I compare the ability of different OTU clustering algorithms using the heritability of the resultant OTUs as a measure of quality. This is carried out on the basis that the most biologically accurate clusterings should produce the highest heritability estimates (as they will produce less technical noise). This is an entirely novel approach to technical benchmarking. The results of this chapter inform the methods selected for use in Chapters 6 and 7, and are discussed further in Chapter 8.

This chapter is presented in the form of a manuscript that was originally published in the journal *PeerJ* on the 30th of August 2016. Accompanying supplementary materials can be found in Appendix C. The digital appendix also includes the original PDF file for the manuscript.

Collaborator Attributions

Collaborators from Cornell University carried out the sequencing and demultiplexing of the 16S rRNA gene reads used in this chapter.

A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units

Matthew A. Jackson, Jordana T. Bell, Tim D. Spector and Claire J. Steves

Department of Twin Research & Genetic Epidemiology, King's College London, University of London, London, United Kingdom

ABSTRACT

A variety of methods are available to collapse 16S rRNA gene sequencing reads to the operational taxonomic units (OTUs) used in microbiome analyses. A number of studies have aimed to compare the quality of the resulting OTUs. However, in the absence of a standard method to define and enumerate the different taxa within a microbial community, existing comparisons have been unable to compare the ability of clustering methods to generate units that accurately represent functional taxonomic segregation. We have previously demonstrated heritability of the microbiome and we propose this as a measure of each methods' ability to generate OTUs representing biologically relevant units. Our approach assumes that OTUs that best represent the functional units interacting with the hosts' properties will produce the highest heritability estimates. Using 1,750 unselected individuals from the TwinsUK cohort, we compared 11 approaches to OTU clustering in heritability analyses. We find that de novo clustering methods produce more heritable OTUs than reference based approaches, with VSEARCH and SUMACLUST performing well. We also show that differences resulting from each clustering method are minimal once reads are collapsed by taxonomic assignment, although sample diversity estimates are clearly influenced by OTU clustering approach. These results should help the selection of sequence clustering methods in future microbiome studies, particularly for studies of human host-microbiome interactions.

Subjects Bioinformatics, Ecology, Microbiology
Keywords Ecology, Microbiology, Computational biology

INTRODUCTION

The field of microbiome research has seen rapid expansion this last decade (*Jones, 2013*). One of the techniques most frequently used to profile microbial communities is 16S rRNA gene sequencing, where PCR amplification of variable marker regions is used to determine a sample's microbial composition (*Pace, 1997*). The taxonomic resolution of sequence variation across a marker region is limited both biologically and technically, because sequence divergence may not represent wider biological divergence between taxa (*Stackebrandt & Goebel, 1994*; *Mignard & Flandrois, 2006*), and sequencing errors introduce artificial divergence (*Huse et al., 2010*; *Schloss, Gevers & Westcott, 2011*). As a result, it is not necessarily useful to enumerate every unique sequence observed particularly

Submitted 7 June 2016
Accepted 18 July 2016
Published 30 August 2016

Corresponding author
Matthew A. Jackson,
matthew.jackson@kcl.ac.uk

Academic editor
Jun Chen

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj.2341**

How to cite this article Jackson et al. (2016), A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. **PeerJ 4:e2341; DOI 10.7717/peerj.2341**

given that samples may contain hundreds of thousands of unique reads. To simplify analyses, reads within a 16S rRNA gene dataset are typically collapsed into operational taxonomic units (OTUs). This is carried out based on sequence similarity between reads. Convention is typically to group reads that share at least 97% identity, which is considered "species" level. Although collapsing can be carried out to any threshold and there is no clear definition of what constitutes a bacterial species.

A variety of methods are available to collapse 16S data to OTUs (*Edgar, 2010*; *Edgar, 2013*; *Rognes et al., 2016*; *Mercier et al., 2013*; *Mahé et al., 2014*; *Schloss & Handelsman, 2005*; *Eren et al., 2014*), often implemented within software wrappers such as QIIME and Mothur (*Caporaso et al., 2010*; *Schloss et al., 2009*). One of the main divides in approaches is whether experimental sequences are clustered against a reference database of sequences (*Liu et al., 2008*), termed closed reference clustering (*Navas-Molina et al., 2013*), or solely clustered within the experimental data itself, generating what are termed de novo OTUs (*Schloss & Handelsman, 2005*; *Navas-Molina et al., 2013*). Closed reference clustering is computationally more efficient given that each sequence should maximally only be compared against each reference sequence, whereas de novo clustering could require pair-wise comparisons between all experimental reads. Closed reference approaches also facilitate comparisons between datasets as OTUs can be defined and matched based on their reference sequences; however, reads which do not match any reference sequences will be discarded. De novo clustering does not have this limitation and includes all experimental reads in resultant OTUs, which may better represent rare and novel taxa (*Navas-Molina et al., 2013*). A third approach, termed open-reference clustering, aims to capitalise on the benefits of both approaches by first clustering experimental sequences against a reference followed by de novo clustering of discarded sequences (*Navas-Molina et al., 2013*).

Once a reference or de novo based approach has been selected, a number of different algorithms can be used to cluster sequences by similarity (*Schloss & Handelsman, 2005*; *Caporaso et al., 2010*; *Edgar, 2010*; *Edgar, 2013*; *Rognes et al., 2016*; *Mercier et al., 2013*; *Mahé et al., 2014*; *Eren et al., 2014*). Linkage based methods calculate pairwise distances between all sequences allowing hierarchical clustering to OTUs (*Schloss & Handelsman, 2005*). There are also multiple greedy algorithms available, which aim to reduce computation time using heuristic approaches to finding optimal groups without calculating all possible distances (*Edgar, 2010*; *Edgar, 2013*; *Rognes et al., 2016*). Furthermore, there have been a number of methods proposed to summarise 16S data without using a predetermined global similarity threshold. These include simply using de-replicated sequences (reads collapsed by 100% similarity), defining OTUs by inherent separation within the dataset using local rather than global cut-offs (*Mahé et al., 2014*), and splitting reads into groups based on sequence entropy at each position in aligned reads (*Eren et al., 2014*).

With the range of available approaches to OTU picking some comparative metric is required to assess their performance. Previously, clustering algorithms have been compared based on a number of metrics including: their computational efficiency (*Edgar, 2010*; *Kopylova et al., 2016*; *Chen et al., 2013*); the number of OTUs they produce (*Schmidt, Rodrigues & Von Mering, 2015*; *Kopylova et al., 2016*; *Chen et al., 2013*); the accuracy of the

similarity between sequences within their OTUs (*Westcott & Schloss, 2015*; *Schloss, Gevers & Westcott, 2011*; *Schloss, 2016*); their ability to handle sequencing artefacts (*Edgar, 2013*); their reconstruction of simulated data sets (*Kopylova et al., 2016*; *Chen et al., 2013*); the similarity between method outputs (*Schmidt, Rodrigues & Von Mering, 2015*; *Kopylova et al., 2016*); and the reproducibility of their clustering within subsets of the same data (*He et al., 2015*). However, the optimal approach between de novo and reference clustering, and the different clustering algorithms is dependent on which measure of quality is considered.

As there is no accepted standard for definition and enumeration of microbial taxa in a community, existing comparison metrics have exclusively dealt with technical aspects of clustering. It is not clear which of these metrics is most important in determining a methods ability to generate OTUs most representative of the biological units underlying microbial community structure. Here we suggest heritability as a measure of the biological relevance of OTUs.

Heritability quantifies the percentage of phenotypic variation that is attributable to genetic variability. Twin studies are a well-established method for estimating heritability. These compare the correlation of phenotypes within monozygotic (MZ) twin pairs whom share identical nuclear DNA, to the correlations within dizygotic (DZ) pairs whom on average share half their genetic material. Variation in a phenotype can then be apportioned into variation due to genetic factors, which are shared by twins to a varying degree, based on zygosity and to environmental factors, which are not shared by twins (*Franic et al., 2012*; *Boomsma, Busjahn & Peltonen, 2002*).

TwinsUK is a long established cohort of unselected British twins (*Moayyeri et al., 2013*). 16S rRNA gene sequencing of faecal samples from the cohort has been used to demonstrate heritability of the microbiome (*Goodrich et al., 2014*; *Goodrich et al., 2016*), and to identify a number of phenotype-microbiome associations (*Jackson et al., 2016a*; *Jackson et al., 2016b*; *Barrios et al., 2015*). Under the assumption that some heritability within the microbiome is acting at the level of individual taxa-host interactions, we propose that the heritability of OTUs is representative of their ability to summarise the underlying biological units within a microbial community.

Here we compare heritability estimates of 11 different methods of summarising 16S reads from 1,750 faecal samples of 473 MZ and 402 DZ twin pairs. Overall, we find that de novo clustering, regardless of algorithm, consistently produces more heritable OTUs than reference based approaches, with VSEARCH and SUMACLUST producing the highest heritability estimates from those considered. No difference in heritability was observed once OTUs had been collapsed by taxonomic assignment. We also find that clustering method can influence relative sample diversity, dependant on the diversity metric used. These results should provide guidance to researchers in selecting the appropriate approach to OTU picking, in particular in studies investigating human host-microbiome interactions.

METHODS

Faecal sampling and 16S rRNA gene sequencing

Analyses were carried out using 16S rRNA gene sequencing reads from a subset of published data from the TwinsUK cohort. Sample collection, DNA extraction and sequencing have

previously been reported (*Goodrich et al., 2014*). In brief, twins produced the sample at home, which was then kept refrigerated and/or on ice before freezing at −80 °C in the TwinsUK laboratory at King's College London. Frozen samples were then shipped to Cornell University where extracted DNA from samples was PCR amplified over the V4 variable region of the 16S gene. The resulting amplicons were multiplexed and sequenced using the Illumina MiSeq platform to generate 250 bp paired-end reads. Ethical approval for microbiota studies within TwinsUK were provided by the NRES Committee London—Westminster (REC Reference No.: EC04/015). All participants provided written consent.

Pre-processing of sequencing reads

Paired reads were joined using fastqjoin, within QIIME (*Caporaso et al., 2010*), discarding reads without a minimum overlap of 200 nt and those containing ambiguous bases. Joined reads were de-multiplexed also removing barcodes. The data were filtered to only include the subset of 1,750 samples from the 473 MZ and 402 DZ complete twin pairs used in these analyses. Within this set, there were 158,635,772 reads with an average of 91,170 reads per sample. These were split per sample and de novo chimera checking carried out on each individually using USEARCH de novo chimera detection in QIIME with a no vote weight of 7 (*Edgar et al., 2011*; *He et al., 2015*). This identified an average of 8,471 chimeric reads per sample all of which were removed. Sample reads were then concatenated to one file and all sequences <252 nt or >253 nt in length discarded (<1% of reads) (*Kozich et al., 2013*). After chimera removal and length filtering, the final data set contained 142,307,280 reads across all samples. This fasta file was used as the input for all 16S collapsing approaches.

These reads and associated metadata, covering a larger selection of samples and twins than the subset described here, are available from the European Nucleotide Archive (ENA) from the study with accession number ERP015317 (*Goodrich et al., 2016*).

Clustering of 16S rRNA gene sequencing reads

All threshold based OTU clustering approaches and Swarm were implemented using QIIME 1.9.0 (*Caporaso et al., 2010*; *Mahé et al., 2014*). VSEARCH de novo clustering was implemented within the QIIME wrappers using an alias to run VSERARCH in place of USEARCH (*Rognes et al., 2016*; *Edgar, 2010*). VSEARCH is not restricted to the same memory limitations as the free version of USEARCH, enabling its use across our whole data set. It also accepts the same commands for de novo clustering so required no alterations to the QIIME wrapper. Where a reference was required, the Greengenes reference and taxonomy version 13_8 was used (*DeSantis et al., 2006*). De-replicated sequences were generated using VSEARCH (*Caporaso, 2015*). Minimum entropy decomposition (MED) was run from scripts within the oligotyping pipeline using default parameters (*Eren et al., 2014*; *Eren et al., 2013*). An overview of how each clustering method works, the clustering pipeline, and complete commands used for each clustering procedure can be found in Supplemental Information 1.

Heritability analyses

Heritability of microbiome traits was calculated in a manner similar to as previously reported (*Goodrich et al., 2014*). Estimates were calculated for OTUs found in at least 50%

of samples as OTU absence, which skews the distribution of abundances, would be less influential on model fitting. A pseudo count of 1 was added to all OTUs to remove absent data in the resultant OTU tables of each clustering approach. Counts were converted to within sample relative abundances and tables subset to only include OTUs found in at least 50% of samples (prior to the addition of pseudo counts). The powerTransform package in R was used to estimate a Box–Cox transform lambda producing approximately normally distributed residuals from a linear model with OTU abundance as a response and gender, age, sequencing run, sequencing depth, how the sample was collected, and the technician who loaded and extracted the DNA as predictors. This was carried out for each OTU and the transformed residuals used in heritability estimation.

Estimates were found by fitting OTU abundances to a twin-based ACE model. This estimates narrow-sense heritability (the heritability due to additive genetic effects—A) on the assumption that variance resulting from shared environment (common environment—C) is equal in MZ and DZ twins, with remaining variance attributed to environmental influences unique to individuals (E) (*Franic et al., 2012*). Maximum likelihood estimates were found by structural equation modelling using OpenMX in R (*R Development Core Team, 2009*; *Boker et al., 2011*). Heritability estimates for collapsed taxonomic traits were calculated in the same manner as for OTUs.

Between method comparisons of OTU heritability and other distributions were carried out in R using pairwise Mann–Whitney *U* tests using Benjamini–Hochberg FDR correction to account for multiple testing.

Alpha diversity calculation and taxonomic assignment

Each complete OTU table was rarefied to 10,000 sequences 25 times. Alpha diversity calculation was carried out on each rarefied table for each method using Simpson, Shannon, Chao1 and raw OTU count metrics, with final diversity values taken as the mean across all rarefactions. Alpha diversity estimates were compared using Mann–Whitney *U* tests to contrast absolute values between methods and Kendall rank correlations to compare sample rankings between methods.

For each clustering method, except closed reference, representative sequences were selected as the most abundant read within each OTU. These were then used to assign taxonomy against the Greengenes 13_8 database with a 97% similarity threshold using the UCLUST method in the assign taxonomy script of QIIME. OTU tables were collapsed based on taxonomic assignment at all levels from genus to phylum. Differences in heritability of taxa between methods were compared using a generalised linear model in R, to determine the ability of taxonomic assignment and clustering method to predict heritability estimates as the response variable. This was carried out across all taxonomic levels considering all taxa that were found across all 11 clustering approaches.

RESULTS

De novo clustering produces more heritable OTUs than closed reference clustering

16S microbiome profiles were available for 473 MZ and 402 DZ pairs within previously reported data. Joined paired end read data were revisited and chimeric sequences removed

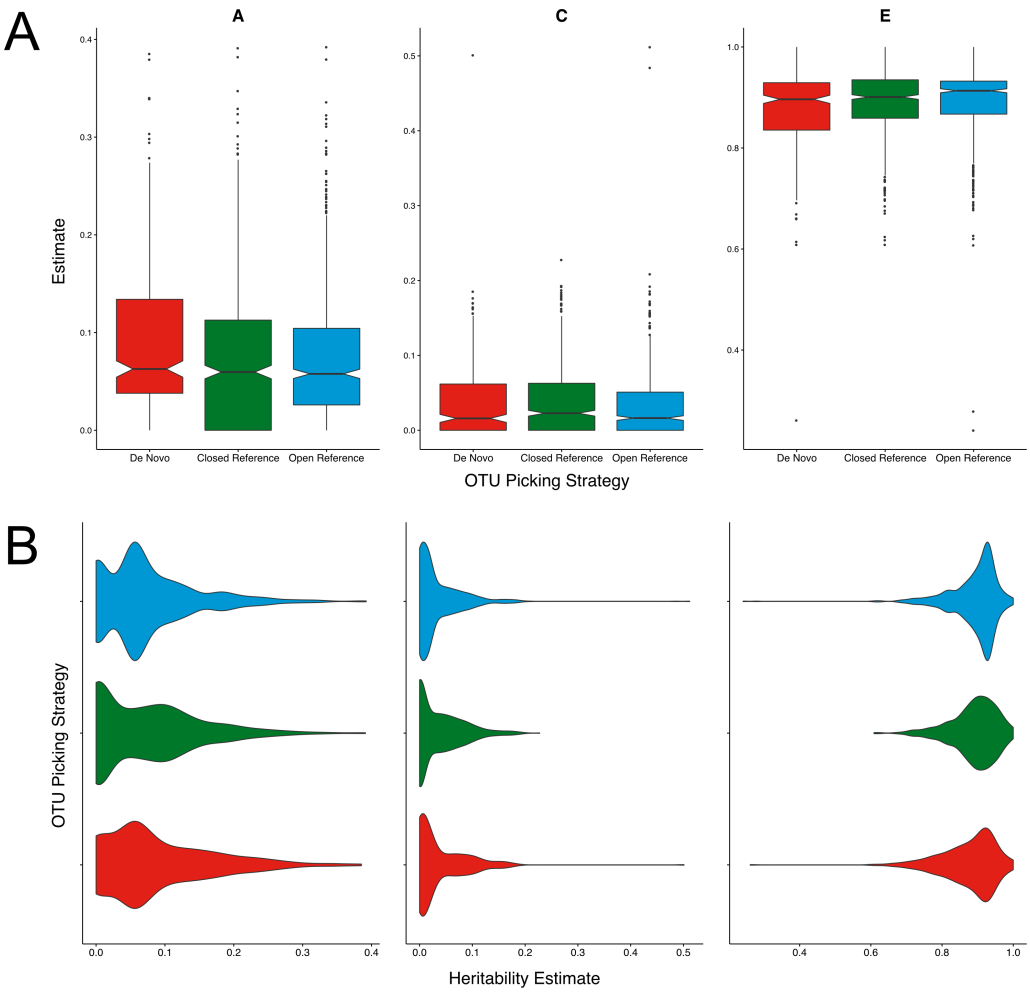


Figure 1 Twin based A, C, and E estimate comparisons between closed and open reference, and de novo clustering using UCLUST with a similarity threshold of 97%. (A) Boxplots representing the A, C and E estimates for all OTUs found in at least 50% of samples in each method. De novo clustering A estimates significantly higher than those of closed reference clustering ($q = 0.017$). (B) The same estimates as in A but displayed as a density function showing the distribution of estimates amongst OTUs.

on a per sample basis. Total read data across all 1,750 samples was then clustered using de novo, closed reference, and open reference approaches using the UCLUST algorithm (*Edgar, 2010*), the current default in QIIME, to form OTUs with a threshold similarity of 97%. The resultant OTU tables are summarised in Table S1. De novo clustering produced more OTUs than closed reference and as a result, a more sparsely distributed OTU table. Open reference picking was an intermediate of the two approaches as might be expected.

Across all three methods the A, C, and E estimates were within the range expected from previous reports within the cohort (*Goodrich et al., 2014*; *Goodrich et al., 2016*). De novo clustering produced OTUs with significantly higher ($q = 0.017$) heritability (A) estimates than closed reference clustering (Fig. 1A). De novo heritability estimates were also higher than those of open reference OTUs although the difference was non-significant. There

were no significant differences in the distributions of C estimates between any methods. De novo clustering produced OTUs with significantly lower E estimates than both closed ($q = 0.02$) and open reference ($q = 0.003$) approaches.

Whilst significant, the difference in OTU heritability estimates was only moderate. The mean of the de novo A estimates was 1% higher than that of the closed reference clustered OTUs. However, the distribution of A, C, and E estimates were also divergent, as shown in Fig. 1B. Closed reference A estimates displayed a bimodal distribution with OTUs either having no or little heritability with fewer highly heritable units. De novo clustering produced units of higher heritability whose estimates were more evenly distributed. Open reference clustering displayed features of both distributions resulting in higher levels of moderately heritable OTUs.

VSEARCH and SUMACLUST produce more heritable de novo OTUs than UCLUST

As de novo clustering produced the most heritable OTUs using UCLUST, we aimed to determine the influence of using alternative threshold based algorithms for clustering. Linkage based clustering approaches were not considered as it was unfeasible to generate distance matrices between the large number of unique reads within the data set. OTUs were clustered at 97% similarity using two alternate greedy algorithms within QIIME—VSEARCH and SUMACLUST (*Rognes et al., 2016*; *Mercier et al., 2013*). The open-source algorithm VSEARCH was used in place of the QIIME default USEARCH to overcome the memory limitations of its free version. VSEARCH has previously been shown to match or outperform USEARCH in terms of accuracy (*Westcott & Schloss, 2015*). Clustering with VSEARCH was carried out using both distance and abundance options as tiebreak assignments. The resultant OTU tables are summarised in Table S1.

There were no significant differences in the mean magnitudes of the A, C, or E estimates between all four methods tested (Fig. 2A). The distributions of estimates were very similar in the SUMACLUST, and both VSEARCH approaches (Fig. 2B). UCLUST OTUs contained a higher proportion of A estimates falling between 0.05 and 0.15, with the other methods containing higher proportions of more heritable OTUs. The VSEARCH methods had more OTUs with high heritability estimates (0.35–0.4), with the distance tiebreaker based method producing slightly fewer. SUMACLUST produced the most heritable OTU. Overall, all de novo algorithms produced estimates higher than the UCLUST reference based approaches at a threshold of 97% similarity, with SUMACLUST and VSEARCH approaches producing more heritable OTUs than UCLUST.

Clustering at higher thresholds and other alternatives to clustering

We aimed to investigate the use of more stringent thresholds repeating VSEARCH abundance based clustering with identity thresholds of 98 and 99%, and simply de-replicating the sequences, the equivalent of a 100% threshold. We also clustered sequences using two approaches that do not rely on a sequence identity threshold—MED and Swarm (described in Supplemental Information 1) (*Eren et al., 2014*; *Mahé et al., 2014*). Of the thresholds, 97% produced the most heritable OTUs (Fig. 3A), whose distribution of A

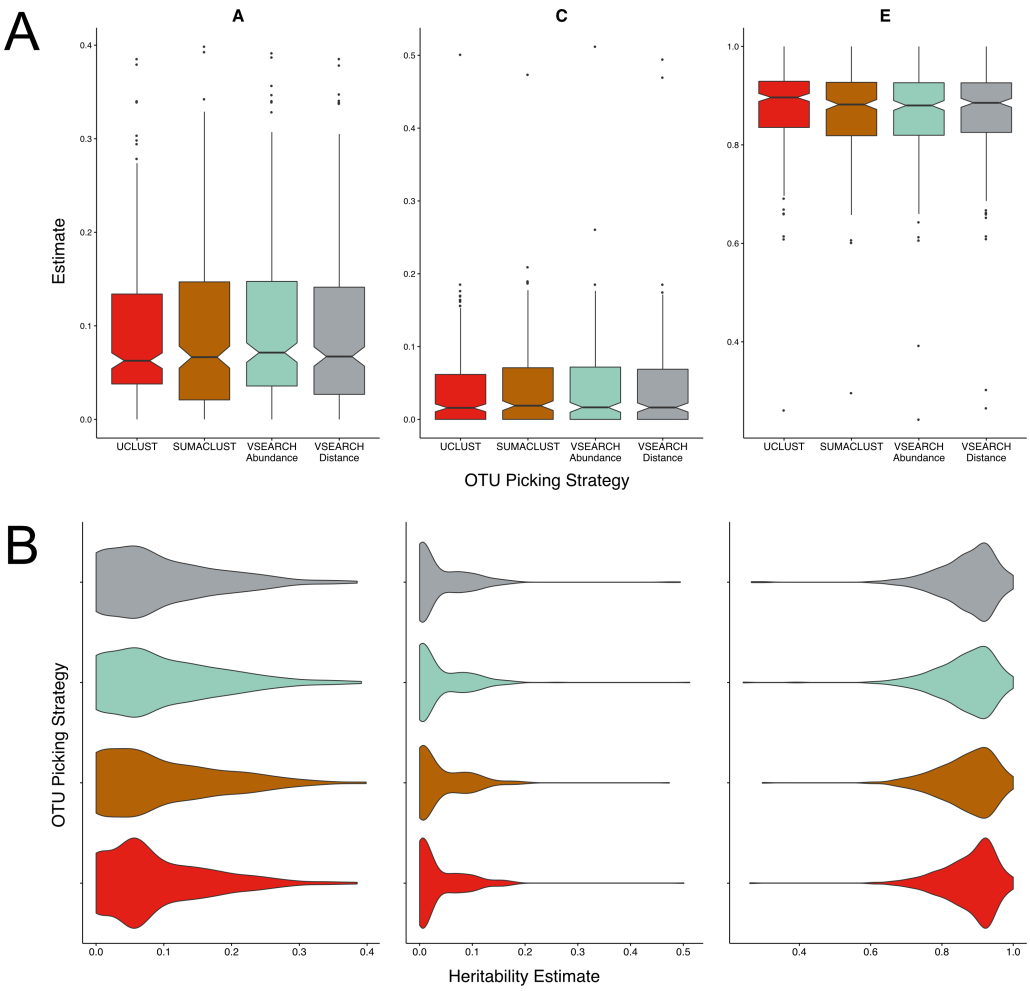


Figure 2 Twin based A, C, and E estimate comparisons between different greedy algorithms for de novo clustering at a 97% similarity threshold. (A) Boxplots representing the A, C and E estimates for all OTUs found in at least 50% of samples in each method. There was no significant difference in A estimates between methods. (B) The same estimates as in A but displayed as a density function showing the distribution of estimates amongst OTUs.

estimates was significantly different to those of the 99 ($q = 0.02$) and 100% ($q = 0.0001$) cut-off OTUs (Fig. 3B). As the percentage identity increased from 97% through to 100% the distribution of A estimates became less continuous, with small groups of units with high heritability and much larger numbers with low heritability. This suggests that in some instances, the heritability estimate of an OTU clustered at 97% identity may be driven by an individual, highly heritable sequence; as opposed to the accumulative effects of the variance across all its reads.

MED produced very few units in total (Table S1). However given this broad level of summary, which is comparable to that of closed reference clustering, the resultant units A estimates were not significantly different to VSEARCH OTUs clustered at the 97% level. Similarly, the heritability of OTUs resulting from clustering by Swarm had heritability's

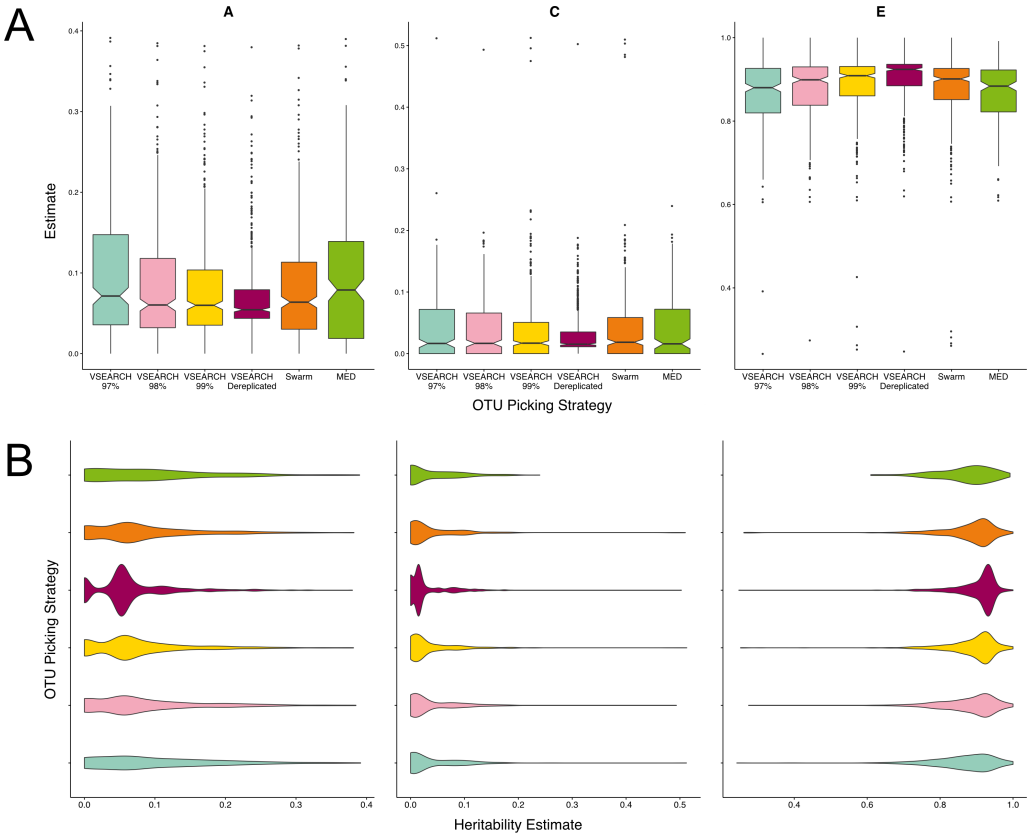


Figure 3 Twin based A, C, and E estimate comparisons between three different thresholds of de novo clustering using VSEARCH, VSEARCH de-replicated sequences, and two non-threshold based techniques. (A) Boxplots representing the A, C and E estimates for all OTUs found in at least 50% of samples in each method. The 97% threshold produced significantly more higher A estimates than the 99 and 100% thresholds ($q = 0.02$, $q = 0.0001$). (B) The same estimates as in A but displayed as a density function showing the distribution of estimates amongst OTUs.

within the range of the VSEARCH methods, however the distribution of A estimates more closely resembled OTU clustering at a threshold of 99%.

De novo clustering at 97% generates more heritable OTUs than reference-based approaches when considering only heritable units

The power of a twin study to detect and accurately estimate the additive genetic variance of a trait is limited by the total number of pairs and the proportion of MZ twins considered (*Visscher, 2004*). As noise in the A estimates for non and low heritability traits may influence the overall distribution, we compared A estimate distributions across all previously clustered techniques considering only heritable OTUs—those with A estimates greater than the mean of all OTUs (8%) and with a lower 95% confidence interval of at least 1% (Fig. 4). When only considering the most heritable OTUs, the majority of de novo based approaches produced units with higher heritability estimates than the reference-based approaches. VSEARCH AGC clustering at 97 and 98%, and DGC clustering at 97% produced significantly higher estimates than closed reference UCLUST. As did SUMACLUST de novo clustering (97%

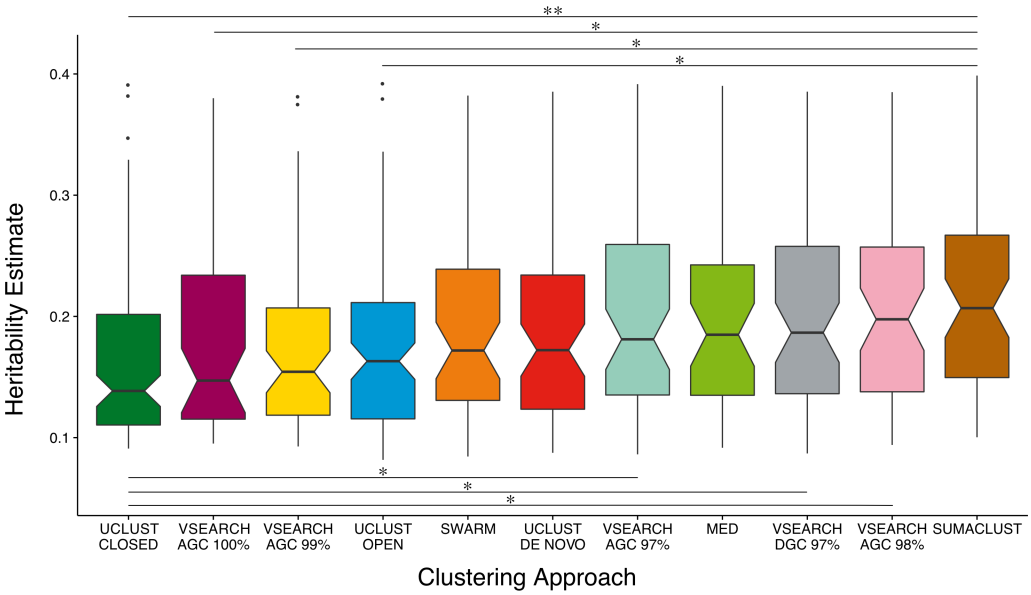


Figure 4 Comparison of A heritability estimates between all clustering approaches. Only considering OTUs who's A estimate was greater than the mean (~8%) and had a lower 95% CI greater than 1%. SUMACLUST and VSEARCH clustering produced OTUs with significantly higher heritability estimates than OTUs produced using reference-based clustering. Significant differences are shown where * indicates $q < 0.05$ and ** indicates $q < 0.01$.

identity), which also produced units with significantly higher heritability than those produced by open reference based clustering. De novo clustering at higher sequence identity thresholds (99 and 100%) produced OTUs with significantly lower estimates than SUMACLUST at 97%.

Differences resulting from clustering approach are not apparent after collapsing by taxonomic assignment

The ability of a technique to generate OTUs representing fine scale biological units may be less important for studies aiming to identify effects at higher taxonomic levels. To determine if choice of OTU clustering approach significantly effected the ability to generate representative taxa we collapsed each OTU table at all taxonomic levels from genus to phylum, and estimated the heritability of taxa at each level (Table S2). We then investigated the ability of taxonomic assignments and clustering methods to predict taxa heritability estimates. We found that assignments to 150 of the 168 taxa found across all 11 methods were significant predictors of heritability, however none of the clustering methods had a significant effect. This suggests that from genus through to higher-level taxonomic summaries there is sufficient collapsing of reads that the previously observed differences in OTU clustering are not apparent.

Alpha diversity measures are influenced by clustering approach

As the largest difference observed between methods was the number of OTUs generated, we aimed to determine the influence of clustering approach on alpha diversity estimates. The

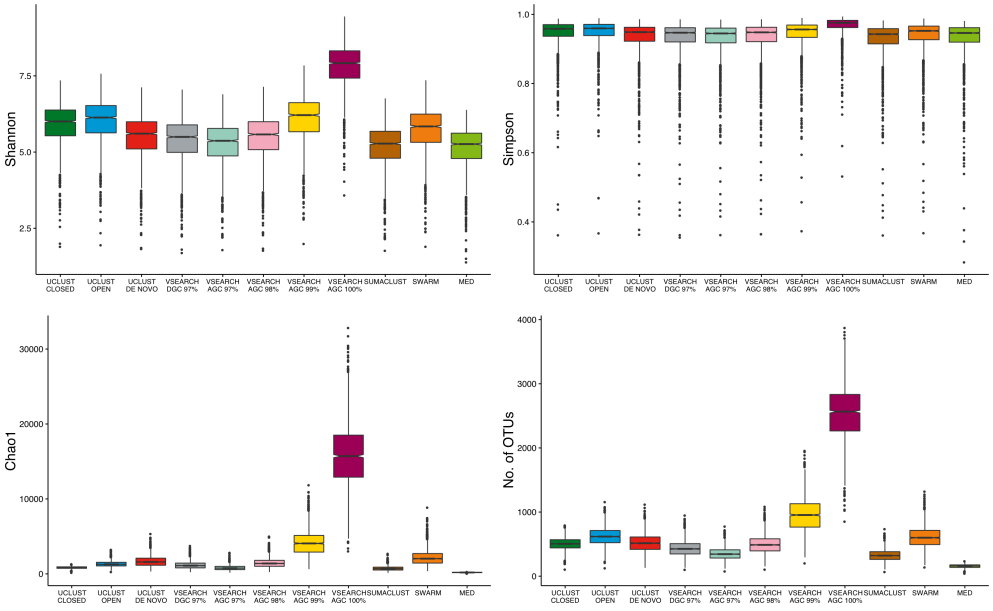


Figure 5 Comparison of absolute alpha diversity values for Shannon, Simpson, Chao1, and OTU count indices across all samples. OTU tables for each method were rarefied to 10,000 sequences 25 times and the mean diversity calculated across all tables. There was a significant difference in the distribution of diversity values between all methods for all four metrics. De-replicated sequences in particular inflate richness-based measures.

absolute values of sample diversity estimates were significantly different between almost all methods of clustering for all four diversity estimates considered (Fig. 5). In particular, the values of OTU count and Chao1 (richness measures influenced by rarer OTUs) were much higher in the de-replicated (or 100% identity) sequences. These results show that absolute diversity levels are not comparable between methods over the same rarefied data.

To determine if these differences would influence comparative diversity analyses, we measured the rank based correlation between methods for each diversity metric (Fig. 6). For both the Shannon and Simpson metrics the diversity rankings were highly correlated ($\tau > 0.6$, mean = 0.83) between all methods. However, when using the Chao1 and OTU count metrics there was a reduced correlation between diversity rankings. In particular, the closed reference and MED approaches were poorly correlated with de novo based approaches. This is likely due to under representation of rare sequences as both of these methods discard reads. Our results show that clustering approach can influence the relative diversities between samples in a study dependant on the diversity measure used. This may be particularly important in the interpretation of diversity association analyses, where use of a closed reference approach could produce different results to the use of de novo clustering.

DISCUSSION

Here we propose and demonstrate the use of heritability estimates as a novel approach to methodological comparisons. There is an established taxa dependent variability in the heritability of the gut microbiome (*Goodrich et al., 2014*). Heritability estimates aim to quantify the percentage of a trait's variation that is due to the influence of host

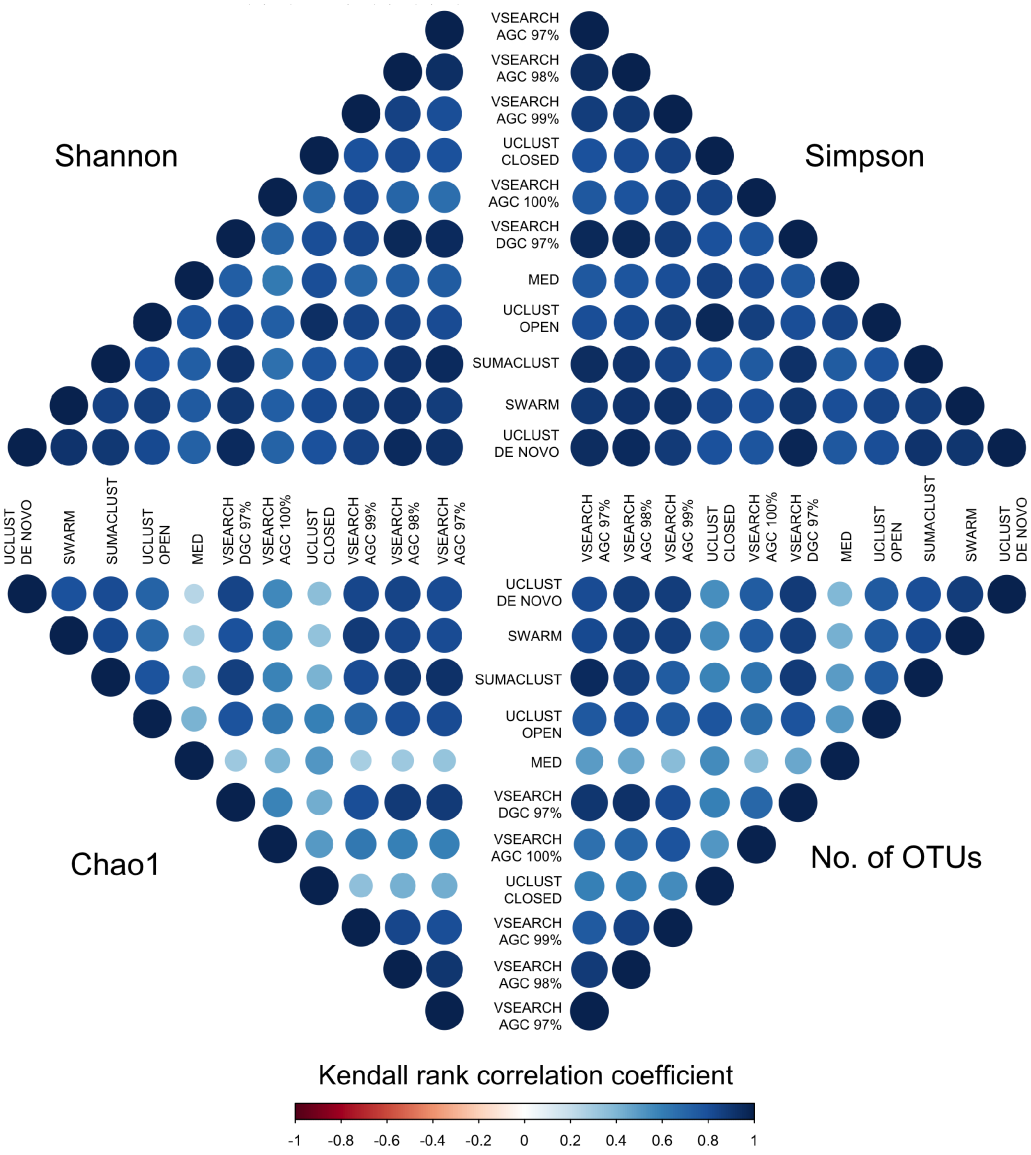


Figure 6 Kendall's Tau rank based correlations between samples across methods for each of Shannon, Simpson, Chao1 and OTU count metrics. Rank correlation represents the concordance between relative diversity assignments between the same samples in each clustering method. There is generally high correlation between all methods when using the Shannon and Simpson indices, which measure evenness of species distribution. However, the de-replicated, closed reference, and MED clustered OTUs show poor correlation in the richness measures (Chao1 and OTU count). Clustering method may therefore influence diversity association analyses.

genetics. Given that bacteria within the microbiome contain a range of functional properties, determined by their own genetics, we assume that the heritability of an OTU is driven by a specific bacteria-host interaction. By this logic, we would expect the OTU clustering approach that best groups reads sourced from bacterial units with similar functional properties to produce OTUs with the highest heritability estimates.

Using the distribution of heritability estimates as a measure of biological representation, we have demonstrated that de novo clustering produces OTUs that are more representative of functional microbial units than reference based approaches. We have also shown that within the various algorithms considered VSEARCH and SUMACLUST produced the most representative OTUs. Within our comparison of clustering thresholds, we found that 97% sequence identity produced the most heritable units when compared to more stringent cut-offs. We have shown that these effects are only applicable at the OTU level, as clustering approach does not significantly influence the heritability estimates of collapsed taxonomies. Finally, we have demonstrated that choice of clustering approach can effect both absolute and relative diversity measures with implications for comparisons across microbial studies.

The aim of OTU clustering is to group sequences based on sequence similarity. Our comparisons are based on the assumption that the genetic relatedness between 16S reads is related to the functional similarity between their bacterial sources. In this way, a clustering method that best groups reads with similar sequence will also groups reads from bacteria with similar functional relationships to the host. These methods should therefore produce the highest heritability estimates, as they will produce less noise in the variance of OTU abundances due to incorrectly grouped read counts. Whilst this may not provide an accurate quantification of the quality of sequence identity within OTUs (as provided by existing methods discussed below), it does provide a measure of the functional representation of the units. For example, in our data the OTUs clustered with 99 and 100% identity thresholds produced lower heritability estimates. Suggesting that 97% is the best threshold to generate units that represent functional units within the microbiome. A methods ability to represent functional units is arguably of more importance than genetic accuracy, particularly for studies in areas such as human microbiome research where the goal is often to identify the functional roles of microbes in human health.

Recently, four studies were published that each compared multiple OTU clustering approaches (*He et al., 2015*; *Kopylova et al., 2016*; *Westcott & Schloss, 2015*; *Schloss, 2016*). The first used the stability of sequence assignments within subsets of the same data sets as a measure of quality, finding that reference based approaches outperformed de novo clustering (*He et al., 2015*). The heritability comparisons presented here do not reflect these findings, suggesting that stability does not relate to functional representation. However, stability may be an important consideration for studies comparing across data sets. Our findings also suggest that reference based approaches would be sufficient when analyses are only concerned with collapsed taxonomies.

Two studies have compared clustering methods using Matthew's correlation coefficient (MCC) to quantify their accuracy in clustering sequences sharing 97% sequence identity (*Westcott & Schloss, 2015*; *Schloss, 2016*). They found that de novo clustering produced more accurate OTUs than reference based approaches (*Westcott & Schloss, 2015*), and that VSEARCH and SUMACLUST out performed Swarm in terms of OTU accuracy (*Schloss, 2016*). The differences between reference and de novo OTUs in our heritability estimates, whilst moderate, were significant and broadly agreed with these observations. This suggests that accuracy is also representative of the biological representation of OTUs. This might be expected under the assumption that sequence similarity, at least in part, reflects functional similarity.

Kopylova et al. (2016) compared a number of clustering methods using a variety of measures from recapitulation of simulated data to inter-method correlations. Within the methods considered here, they found that Swarm, SUMACLUST and UCLUST, performed equally well at reconstructing expected taxonomies from simulated data but differed in the number of OTUs produced and subsequently produced different absolute diversities, a finding also described by *Schmidt, Rodrigues & Von Mering (2015)*. Differences in absolute measures would be expected given the variation in OTU numbers between methods. We have also shown that these differences can influence the relative diversity rankings between samples and suggest caution in the interpretation of comparative diversity analyses when using closed reference clustering and community richness metrics.

Overall, across previous comparisons of greedy clustering algorithms in combination with the heritability results we have presented here, VSEARCH and SUMACLUST seem to produce the best combination of accuracy, stability and heritability. We would therefore recommend either of these approaches for de novo clustering. SUMACLUST and USEARCH are currently available within QIIME. VSEARCH has recently been implemented within Mothur (*Westcott, 2016*), and QIIME 2 will integrate VSEARCH for OTU clustering and de-replication (Greg Caporaso, personal communication, 15th April 2016). Based on our threshold comparisons a similarity cut-off of 97% appears optimal, however this threshold may be specific to VSEARCH application to faecal samples as optimal thresholds can vary by the complexity of the microbial communities under investigation and the method used (*Chen et al., 2013*).

Whilst we tried to include the most frequently used approaches, our study is not comprehensive. We restricted the majority of our comparisons to clustering algorithms that were available within the QIIME pipeline; however, even in this respect, our comparison was not exhaustive. There are further reference based clustering algorithms such as BLAST and SortMeRNA that were not considered (*Camacho et al., 2009*; *Kopylova, Noe & Touzet, 2012*), and de novo approaches such as USEARCH and CD-HIT (*Edgar, 2010*; *Li & Godzik, 2006*). We chose to implement clustering via QIIME as it is one of the most widely used methods to generate OTUs and provided stability in other areas of the processing pipeline, such as taxonomic assignment, which improved comparability. However, QIIME does not implement all OTU clustering algorithms and all of those compared here can also be run independently of QIIME, with a number of them having newer versions available that could influence clustering. Our comparison is also limited by the exclusion of linkage-based approaches, as typically implemented using the Mothur pipeline (*Schloss et al., 2009*). These were not considered in our comparison due to the high computational burden of generating the pair-wise sequence distance matrices that these methods require. Computing time and memory limits were met even when applying additional sequence filtering or restricting distance calculation by taxonomy (*Kozich et al., 2013*). Previous MCC accuracy comparisons showed that average based linkage clustering were as or more accurate than the best de novo approaches dependent on the dataset considered (*Schloss, 2016*). Given the reflection between the MCC and heritability results we might speculate that average linkage based approaches could produce biologically relevant units equivalent to the de novo algorithms we considered.

Our comparisons are further limited as we have only considered sequencing from human faecal samples of a single population. A sufficiently large sample is required to determine heritability estimates for moderately heritable traits (*Martin et al., 1978*); however, clustering and analysis of data on this scale is time consuming and computationally intensive, making it non-trivial to incorporate additional data. There are also few twin microbiome data sets available at the scale of TwinsUK. It is known that existing measures of clustering quality can be data set dependent (*Schloss, 2016*; *Chen et al., 2013*; *Kopylova et al., 2016*). Therefore, our results may not be applicable to non-faecal samples. However, they should be of particular relevance when experiments aim to study the functional aspects of the human gut microbiome.

In conclusion, heritability analyses can be used to provide a measure of the quality of the functional representation of OTUs. This may be used for additional guidance in selecting an appropriate clustering approach in combination with the other comparative metrics available, although the optimum method will be largely dependent on each studies experimental and analytical requirements.

ACKNOWLEDGEMENTS

We would like to thank Julia Goodrich, Andrew Clark, and Ruth Ley of the Department of Molecular Biology and Genetics at Cornell University, our collaborators on the collection, processing, and analysis of the TwinsUK 16S gut microbiome data, whom provided guidance and comments on this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The TwinsUK microbiota project was funded the National Institutes of Health (NIH) RO1 DK093595, DP2 OD007444. TwinsUK received funding from the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013), the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. CJS is funded under a grant from the Chronic Disease Research Foundation (CDRF). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Institutes of Health (NIH): RO1 DK093595, DP2 OD007444.
Wellcome Trust; European Community's Seventh Framework Programme: FP7/2007-2013.
National Institute for Health Research (NIHR).
Chronic Disease Research Foundation (CDRF).

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Matthew A. Jackson conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Jordana T. Bell and Tim D. Spector contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Claire J. Steves contributed reagents/materials/analysis tools, reviewed drafts of the paper, supervised all work carried out by Matthew A. Jackson.

Human Ethics

The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):

Ethical approval for microbiota studies within TwinsUK were provided by the NRES Committee London—Westminster (REC Reference No.: EC04/015).

Data Availability

The following information was supplied regarding data availability:

Sequencing data used within these experiments is available as part of the European Nucleotide Archive (ENA) study with the accession number ERP015317.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2341#supplemental-information.

REFERENCES

Barrios C, Beaumont M, Pallister T, Villar J, Goodrich JK, Clark A, Pascual J, Ley RE, Spector TD, Bell JT, Menni C. 2015. Gut-microbiota-metabolite axis in early renal function decline. *PLoS ONE* **10**:e0134311 DOI 10.1371/journal.pone.0134311.

Boker S, Neale M, Maes H, Wilde M, Spiegel M, Brick T, Spies J, Estabrook R, Kenny S, Bates T, Mehta P, Fox J. 2011. OpenMx: an open source extended structural equation modeling framework. *Psychometrika* **76**:306–317 DOI 10.1007/s11336-010-9200-6.

Boomsma D, Busjahn A, Peltonen L. 2002. Classical twin studies and beyond. *Nature Reviews Genetics* **3**:872–882 DOI 10.1038/nrg932.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Caporaso JG. 2015. VSEARCH-based sequence dereplication through generation of a biom table. *Available at https://gist.github.com/gregcaporaso/f3c042e5eb806349fa18* (accessed on 20 April 2016).

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky

JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**:335–336 DOI 10.1038/nmeth.f.303.

Chen W, Zhang CK, Cheng Y, Zhang S, Zhao H. 2013. A comparison of methods for clustering 16S rRNA sequences into OTUs. *PLoS ONE* **8**:e70837 DOI 10.1371/journal.pone.0070837.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72**:5069–5072 DOI 10.1128/AEM.03006-05.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–2461 DOI 10.1093/bioinformatics/btq461.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**:996–998 DOI 10.1038/nmeth.2604.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**:2194–2200 DOI 10.1093/bioinformatics/btr381.

Eren MA, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML. 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution* **4**:1111–1119 DOI 10.1111/2041-210X.12114.

Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. 2014. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal* **9**:968–979 DOI 10.1038/ismej.2014.195.

Franic S, Dolan CV, Boorsboom D, Boomsma DI. 2012. Structural equation modeling in genetics. In: *Handbook of structural equation modeling*. New York: Guilford Press.

Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic determinants of the gut microbiome in UK Twins. *Cell Host & Microbe* **19**:731–743 DOI 10.1016/j.chom.2016.04.017.

Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human genetics shape the gut microbiome. *Cell* **159**:789–799 DOI 10.1016/j.cell.2014.09.053.

He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R, Zhou H-W. 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* **3**: Article 20 DOI 10.1186/s40168-015-0081-x.

Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12**:1889–1898 DOI 10.1111/j.1462-2920.2010.02193.x.

Jackson MA, Goodrich JK, Maxan M-E, Freedberg DE, Abrams JA, Poole AC, Sutter JL, Welter D, Ley RE, Bell JT, Spector TD, Steves CJ. 2016a. Proton pump inhibitors

alter the composition of the gut microbiota. *Gut* **65**:749–756 DOI 10.1136/gutjnl-2015-310861.

Jackson MA, Jeffery IB, Beaumont M, Bell JT, Clark AG, Ley RE, O'Toole PW, Spector TD, Steves CJ. 2016b. Signatures of early frailty in the gut microbiota. *Genome Medicine* **8**: Article 8 DOI 10.1186/s13073-016-0262-7.

Jones S. 2013. Trends in microbiome research. *Nature Biotechnology* **31**:277–277 DOI 10.1038/nbt.2546.

Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou H-W, Rognes T, Caporaso JG, Knight R. 2016. Open-source sequence clustering methods improve the state of the art. *mSystems* **1**:e00003–15.

Kopylova E, Noe L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**:3211–3217 DOI 10.1093/bioinformatics/bts611.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. *Applied and Environmental Microbiology* **79**:5112–5120 DOI 10.1128/AEM.01043-13.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**:1658–1659 DOI 10.1093/bioinformatics/btl158.

Liu Z, DeSantis TZ, Andersen GL, Knight R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research* **36**:e120–e120 DOI 10.1093/nar/gkn491.

Mahé F, Rognes T, Quince C, De Vargas C, Dunthorn M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**:e593 DOI 10.7717/peerj.593.

Martin NG, Eaves LJ, Kearsey MJ, Davies P. 1978. The power of the classical twin study. *Heredity* **40**:97–116 DOI 10.1038/hdy.1978.10.

Mercier C, Boyer F, Bonin A, Coissac E. 2013. SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences. In: *Programs and Abstracts of the SeqBio 2013 workshop (Abstract), GdRBIM and gdrIM, Montpellier, France.* 27–29. *Available at http:// metabarcoding.org/ sumatra*.

Mignard S, Flandrois JP. 2006. 16S rRNA sequencing in routine bacterial identification: a 30-month experiment. *Journal of Microbiological Methods* **67**:574–581 DOI 10.1016/j.mimet.2006.05.009.

Moayyeri A, Hammond CJ, Valdes AM, Spector TD. 2013. Cohort profile: TwinsUK and healthy ageing twin study. *International Journal of Epidemiology* **42**:76–85 DOI 10.1093/ije/dyr207.

Navas-Molina JA, Peralta-Sánchez JM, González A, McMurdie PJ, Vázquez-Baeza Y, Xu Z, Ursell LK, Lauber C, Zhou H, Song SJ, Huntley J, Ackermann GL, Berg-Lyons D, Holmes S, Caporaso JG, Knight R. 2013. Advancing our understanding of the human microbiome using QIIME. *Methods in Enzymology* **531**:371–444 DOI 10.1016/B978-0-12-407863-5.00019-8.

Pace NR. 1997. A molecular view of microbial diversity and the biosphere. *Science* **276**:734–740 DOI 10.1126/science.276.5313.734.

R Development Core Team. 2009. Vienna: R Foundation for Statistical Computing.

Rognes T, Mahé F, Flouri T, Quince C, Nichols B. 2016. VSEARCH. *Available at https://github.com/torognes/vsearch*.

Schloss PD. 2016. Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods. *mSystems* **1**:e00027–16 DOI 10.1128/mSystems.00027-16.

Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6**:e27310 DOI 10.1371/journal.pone.0027310.

Schloss PD, Handelsman J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71**:1501–1506 DOI 10.1128/AEM.71.3.1501-1506.2005.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**:7537–7541 DOI 10.1128/AEM.01541-09.

Schmidt TSB, Matias Rodrigues JF, Von Mering C. 2015. Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environmental Microbiology* **17**:1689–1706 DOI 10.1111/1462-2920.12610.

Stackebrandt E, Goebel BM. 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic and Evolutionary Microbiology* **44**:846–849 DOI 10.1099/00207713-44-4-846.

Visscher PM. 2004. Power of the classical twin design revisited. *Twin Research* **7**:505–512 DOI 10.1375/1369052042335250.

Westcott SL. 2016. Mothur. Version 1.37.0. *Available at https://github.com/mothur/mothur/releases/tag/v1.37.0* (accessed on 20 April 2016).

Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**:e1487 DOI 10.7717/peerj.1487.

Chapter 6

Identification of taxonomic markers that define a health-associated gut microbiome

In the following chapter I present an association study considering multiple disorders and medications within the same cohort. This is achieved using a novel method to determine marker taxa in the microbiota. I then assign these markers as generally more health or disorder associated based on the observed associations. The ratio of these two groups provides a novel definition of a health-associated gut microbiome. I then quantify this definition in an index called the health-associated microbiome index (HMI) and demonstrate its use in gut microbiota research. The microbiota markers and HMI address the high-dimensionality and limited reproducibility of 16S rRNA gene sequencing data and provide a novel clinical tool for gut microbiome research.

Collaborator Attributions

Parsing of disease and drug data from questionnaires was carried out by Serena Verdi, Cheol Min Shin, Maria-Emanuela Maxan, and Claire Steves. Dietary data was analysed by Ruth Bowyer. Metagenomic data was pre-processed by collaborators at BGI as described by Xie et al. (Xie et al. 2016). Faecal metabolomics analyses were carried out by Jonas Zierer.

6.1 Introduction

As discussed in Section 1.2.2, the composition of the human gut microbiome is associated with a wide range of health disorders, from those affecting the gastrointestinal (GI) tract to distal associations with neurological disease, as well as measures of overall health such as frailty (Chapter 3) (Gevers et al. 2014; Sampson et al. 2016; Claesson et al. 2012). However, it has yet to be established if there are differences in the gut microbiome that are shared across disorders. Identification of more widely health-associated features in the gut microbiome could elucidate common microbial functionalities contributing to disease, provide a novel measure of health, and a single clinical target for interventions in multiple diseases.

Comparison of gut microbiome associations across multiple diseases is hindered by the inherent experimental and analytical variability between disease specific studies (see Chapter 1). Reproducibility of microbiome associations to date has often been poor even for a single disorder (Sze and Schloss 2016). These limitations can be ameliorated by comparing associations in a single study with uniform experimental and analytical approaches, as demonstrated by large scale phenotype comparisons in the Flemish Gut and Dutch Lifelines-DEEP cohorts (Falony et al. 2016; Zhernakova et al. 2016). However, these studies focused on determining the host factors most influencing gut microbiome composition rather than commonalities in bacterial associations across diseases.

The longevity of the TwinsUK cohort has lead to the accumulation of a wide range of detailed health phenotype data across individuals for whom gut microbiota profiles are also available. Furthermore, as members of TwinsUK are older than those of other population based cohorts with gut microbiome data (Zhernakova et al. 2016; Falony et al. 2016; Consortium 2012), the cohort should contain more cases of age-related disease. TwinsUK therefore provides a uniform platform to carry out gut microbiota association analyses with multiple common diseases. In this chapter, I carry out gut microbiota association analyses of 38 common health disorders and 51 common prescription medications in TwinsUK. This overcomes the limitations of reproducibility associated with 16S rRNA gene sequencing-based microbiota profiles and enables direct comparisons between the results.

Studying multiple health related traits in tandem dramatically increases the number of statistical tests required to identify associations with the already high-dimensional microbiota data. To overcome this limitation I present a novel method using the inter-correlation between gut microbiota traits to identify marker taxa that are representative of wider microbiota composition.

Comparing associations with the defined marker taxa across the multiple disorders studied, I identify taxa in the gut microbiome that have consistently positive or negative associations with multiple disorders. I define a more health-associated gut microbiome as one with a higher-abundance of consistently health-associated taxa. On this basis, I propose a further method to summarise gut microbiota composition by scoring the 'health-association' of

an individuals' gut microbiota using a health-associated microbiome index (HMI). This provides a novel method to quantify microbiota composition in relation to host health in a single measure. I then demonstrate the potential utility of the HMI by using it to investigate host factors and microbial functions associated with a health-associated microbiome.

6.2 Methods

6.2.1 Disorder and drug data

Self-reported disorder data was collected from 6 questionnaires completed by members of the TwinsUK cohort at various times between 2002-2015. The majority of diseases were scored from the BCQ and Q11A questionnaires, which most twins had answered within 2 years of the faecal samples used to assess the gut microbiota (Figure D.1). All questions asked if a doctor or health professional had ever diagnosed the individual with the condition. Individuals were scored positive for a disorder if they replied yes on any questionnaire, negative if they only replied no, and unknown if data was unavailable across all questionnaires. For constipation and cystitis, questionnaire data was scored as (0) No, (1) Rarely, (2) Sometimes, (3) Frequently, and (4) Always; in these two cases 0-2 was considered negative and 3-4 positive. Hearing loss was classified by either doctor diagnosis, self-diagnosis, or hearing aid usage. For some conditions, for instance food allergies, multiple questions were combined and individuals scored positive if they had reported having any one of the conditions. Status was determined for a comprehensive range of disorders across the wider cohort, and those found in at least 1% of the total cohort were considered common and retained in analyses. The final list of disorders considered is shown in Table 6.1.

Self-reported prescription drug use was available from one of the questionnaires. Reported drug data was initially cleaned computationally to resolve spelling errors, followed by manual classification of entries into drug classes and sub-classes by a health professional. Individuals were assumed not to be taking a drug if they had completed the questionnaire without listing it. Drugs used by at least 1% of the total cohort were considered for further analysis (Table 6.2). Correlations between drugs and between diseases were assessed using the Phi coefficient, equivalent to Pearson's for binary variables.

6.2.2 Processing of 16S rRNA gene sequences

These analyses used the most recent and largest set of gut microbiota samples available within TwinsUK (Batch 3, see Table 2.1). Collection and processing of the faecal samples within this set is consistent with those used throughout this thesis (Section 2.2). Merging of paired-end reads and demultiplexing by sample was carried out by collaborators at Cornell.

Table 6.1: The 38 common disorders (>1% of TwinsUK) considered in this chapter.

Identifier	Description
Gout	
RheumatoidArthritis	
Osteoarthritis	
AMD	Age Related Macular Degeneration
Diverticular Disease	
Coeliac Disease	
Hypertension	
Cholelithiasis	
Hypercholesterolaemia	
Breast Cancer	
Incontinence	
Psoriasis	
Acne	
Cold Sores	
Glaucoma	
Cataract	
COPD	Chronic Obstructive Pulmonary Disease (Chronic Bronchitis, Chronic Obstructive Pulmonary Disease, and Emphysema)
Allergy (Respiratory)	Cats, Dogs, Dust Mites, Pollen and Mould or Mildew
Allergy (Ingested)	Nuts, Penicillin, Fish, Shellfish, Alcohol and Fruits
Allergy (Skin)	Metals and Latex or Rubber
Fracture Risk	
IHD	Ischaemic Heart Disease (Coronary Heart Disease, Myocardial Infarction, Angina and Ischemia)
AF	Atrial Fibrillation
T2DM	Type 2 Diabetes
Depression	
Asthma	
Eczema	
Anxiety	
Constipation	
Hypothyroidism	
Hyperthyroidism	
Recurrent UTI	Recurrent Urinary Tract Infections
IBS	Irritable Bowel Syndrome
Venous Thrombosis	
Epilepsy	
Hearing loss	
Migraine	
IBD	Inflammatory Bowel Disease

Table 6.2: The 51 common medications (>1% of TwinsUK) considered in this chapter.

Identifier	Description
Inhaler - SABA	Inhaler - Beta Agonist - Short-Acting Beta Agonists
Inhaler - LABA	Inhaler - Beta Agonist - Long-Acting Beta Agonists
Inhaler - Steroid	Inhaler - Steroid
Inhaler - Anticholinergic	Inhaler - Anticholinergic
SSRI	Oral - Antidepressant - Selective serotonin reuptake inhibitor
Levothyroxine	Oral - Thyroid Hormone - Levothyroxine
Triptan	Oral - Triptan
Bisphosphonate	Oral - Bisphosphonate
Paracetamol	Oral - Analgesic - Paracetamol
Opioid	Oral - Analgesic - Opiod
Statin	Oral - Statin
Topical - HRT	Topical - Hormone replacement therapy
Calcium	Oral - Calcium
Cholecalciferol	Oral - Cholecalciferol
PPI	Oral - Antacid - Proton Pump Inhibitor
H1 Inhibitor	Oral - Antihistamine - H1 Inhibitor
Quinine	Oral - Antimalarial - Quinine
Antispasmodic	Oral - Antispasmodic
Prescr. Antibiotic	Oral - Antibiotic - Prescription
Contraceptive	Oral - Contraceptive
Beta Blocker	Oral - Antihypertensive - Beta Blocker
ACEi	Oral - Antihypertensive - Angiotensin-converting-enzyme inhibitor
Coumarin	Oral - Anticoagulant - Coumarin
GABAergic	Oral - Analgesic - Gamma-aminobutyric acid system modulator
Thiazide	Oral - Diuretic - Thiazide
Sartan	Oral - Antihypertensive - Sartan
Thyroxine	Oral - Thyroid Hormone - Thyroxine
Topical - Steroid	Topical - Steroid
Steroid	Oral - Steroid
HRT	Oral - Hormone replacement therapy
TCA	Oral - Antidepressant - Tricyclic
CCB	Oral - Antihypertensive - Calcium Channel Blocker
Anticholinergic	Oral - Anticholinergic
Aspirin	Oral - Antiplatelet - Aspirin
Topical - NSAID	Topical - Analgesic - Non-steroidal anti-inflammatory
Benzodiazepine	Oral - Hypnotic - Benzodiazepine
Clopidogrel	Oral - Antiplatelet - Clopidogrel
Laxative	Oral - Laxative
Aromatase Inhibitor	Oral - Hormone Treatment - Aromatase Inhibitor
Metformin	Oral - Anti-diabetic - Metformin
Methotrexate	Oral - Immunosuppressant - Methotrexate
Folic Acid	Oral - Folic Acid
Iron	Oral - Iron
Alginate	Oral - Antacid - Alginate
H2 Antagonist	Oral - Antacid - H2 Antagonist
Loop	Oral - Diuretic - Loop
Alpha Blocker	Oral - Antihypertensive - Alpha Blocker
Vitamin B12	Oral - Vitamin - B12
Injection - Insulin	Injection - Anti-diabetic - Insulin
Glaucoma Eye Drop	Topical - Glaucoma Eye Drop
Aminosalicylate	Oral - Aminosalicylate
Prev. Month Antibiotic	Oral - Antibiotic - Within one month prior to faecal sample

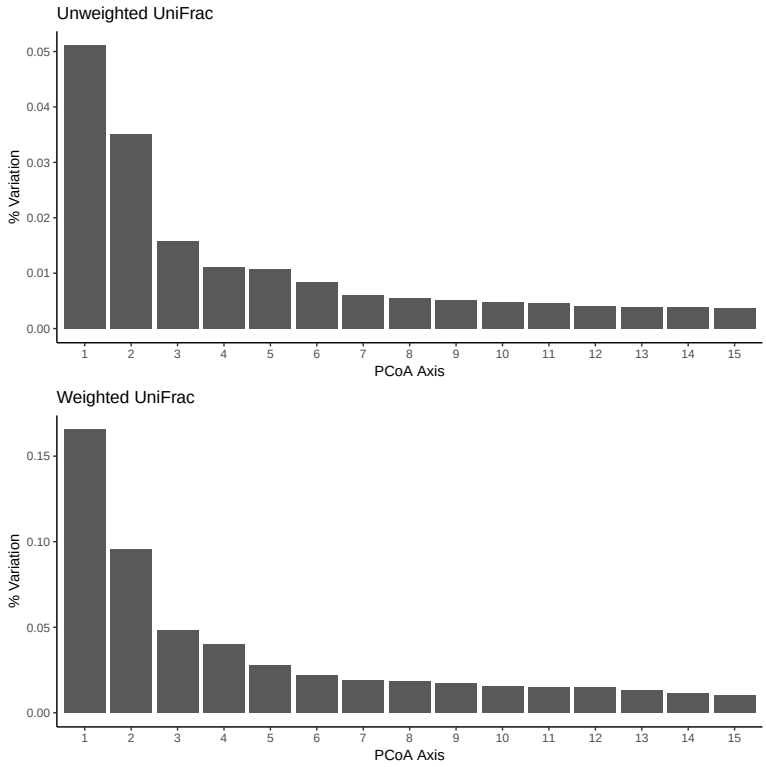


Figure 6.1: Scree plot showing the percentage variation explained by each axis in PCoA of the weighted and unweighted UniFrac distances for all twins in the analysis. The first 6 were considered for later analyses, from both the weighted and unweighted UniFrac PCoAs, as the variance explained plateaued at this point.

These reads were then de novo clustered to OTUs using SUMACLUST with a 97% similarity threshold, following the optimal method described in Chapter 5.

In brief, de novo chimera identification and removal was carried out per sample using UCHIME (Edgar et al. 2011). Remaining reads were then collapsed to OTUs via de novo clustering with SUMACLUST using QIIME version 1.9.0 (Mercier et al. 2013; Navas-Molina et al. 2013). OTU taxonomy was assigned by matching representative sequences against the Greengenes v8_13 database using UCLUST in QIIME (DeSantis et al. 2006).

Taxonomic abundances were generated by collapsing OTU counts at all taxonomic levels, followed by conversion to log transformed relative abundances. Three different alpha diversity metrics, namely the Shannon index, phylogenetic diversity, and raw OTU counts, were calculated using QIIME. Beta diversity was quantified in both weighted and unweighted UniFrac metrics, and principal coordinate analysis (PCoA) of the beta distances was carried out using the vegan package in R (Oksanen et al. 2016). The first 6 axes from each of these PCoAs were chosen to represent beta diversity (Figure 6.1).

6.2.3 Greedy selection of marker traits

As a first step to reduce the dimensionality of the microbiota data a limited selection of traits were chosen for consideration. The three alpha diversity measures, the 12 beta diversity

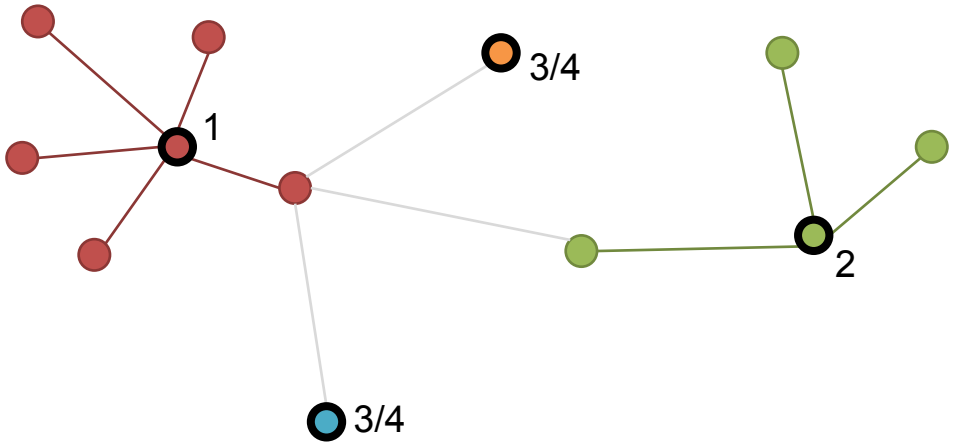


Figure 6.2: Graphical example of the approach used to select marker traits in the gut microbiota. Each node represents a microbiota trait (alpha diversity, beta diversity PC, family, or class abundance) and edges where the Spearman correlation between them is >0.8. Black circles indicate which would be the marker traits and the traits they represent are coloured similarly. Numbers show the order that they are selected. The most connected trait becomes the first marker and all connected nodes are then ignored. This is carried out iteratively until a marker set remains that represents all nodes. In the case of a tie a node is selected at random.

axes, and all collapsed bacterial classes and families with complete taxonomic assignment were chosen as the 206 initial traits representing the microbiota. A novel heuristic approach was then used to select a more limited set of microbiota markers from this set. Spearman correlations were calculated pairwise between the traits and the correlations used to generate an adjacency matrix, in which nodes represented traits and correlations >0.8 represented an edge between them. A graphical representation of the adjacency matrix was then used for greedy selection of representative markers.

Nodes were sorted by degree and the one with highest degree was chosen as a marker trait (selecting at random in the case of a tie). The marker and all connected nodes were then removed from the network and the process repeated until a final set of marker traits was found. This resulted in a set of markers whereby every one of the original 206 traits had a correlation of >0.8 to at least one marker. An example of the approach is visualised in Figure 6.2. After running this selection on the 206 initial microbiota traits considered, a final set of 68 marker traits, containing both alpha and beta diversity measures and class and family abundances, was selected (Table D.1).

6.2.4 Regression analyses

The microbiota marker traits were modelled as responses in mixed effects models with technical and biological confounders including: who extracted the DNA, how the sample was collected, sample sequencing run, sequencing depth, gender, family structure, age, and BMI. The residuals of the variance in the microbiota traits from these models were then used in downstream analyses.

To identify associations between disorders and the microbiota markers, logistic regressions were carried out with disorder status as the dependent variable and the residuals of microbial marker traits as independent variables. This was carried out for all combinations and p-values were FDR corrected to account for multiple testing using the p.adjust command in R. The same approach was also used to investigate drug-microbiome associations. The disorder association analyses were also repeated with residuals created without adjusting for BMI to determine disorder associations additionally considering obesity.

6.2.5 Inference of a Bayesian network between disorders, drug use, and microbiome marker traits

In an attempt to delineate the drug and disease associations with the gut microbiota, I used the CGBayesNet package to infer a Bayesian network structure underlying the microbiome, disorder, and drug use data (McGeachie, Chang, and Weiss 2014). This package was selected as it is designed to incorporate a mix of discrete and continuous variables. As the method requires complete data I selected only individuals who had completed the drug use questionnaire, and assumed that all unknown statuses in the disorder table were negative cases. The residual variances after adjusting for technical and biological covariates were used to represent microbiota marker traits. The command FullBNLearn was used to infer the network using an exhaustive approach whereby iteratively, starting from an empty network, the most likely edge addition or removal is carried out until the posterior likelihood of the network cannot be improved. The algorithm was run using a prior sample size of 10, a prior variance of 1, and allowing a maximum of 3 parents per node. These parameters were described previously in a study using Bayesian networks to predict longitudinal dynamics in the infant gut microbiome (McGeachie et al. 2016). The resultant graph was visualised using Gephi (Bastian, Heymann, and Jacomy 2009).

6.2.6 Clustering of microbiome traits by disorder associations

The beta coefficients of association between the disorders and microbiome traits were filtered to retain only those from nominally significant associations (non-significant association beta coefficients were changed to 0). Microbiome traits and disorders without any significant associations were removed. Distance matrices between disorders and between microbial traits were derived from the beta coefficient matrix using cosine similarity, as it is less influenced by the sparsity resulting from zeroes of non-significant associations. Complete-linkage hierarchical clustering was used to cluster the disorders and microbial traits from the distance matrices, and the resultant clustering visualised as a heatmap. The significance of the resulting microbiome trait clusters was determined by multi-scale bootstrap clustering with 10000 iterations using the pvclust package in R (Suzuki and Shimodaira

2015).

6.2.7 Calculating the HMI and MDI

Following the heatmap clustering I defined an index comparing the ratio of health and disorder associated taxa in the gut microbiome samples. The healthy microbiome index (HMI) was based on a related index defined by Gevers *et al.* - the microbial dysbiosis index (MDI) (Gevers et al. 2014). The HMI and MDI were calculated from counts at the family level as this is the finest taxonomic resolution in the definition of both. Counts were summed into either the numerator or denominator groups within each index based on the familial definitions. Where a higher-level taxonomy was defined, for instance the classes in the HMI, every family within the higher taxa was included into the parent group. In the case where a family and its parent class were assigned to different groups, family assignments overruled class and the counts were added to the family level group. All taxa not in the group definitions were ignored.

6.2.8 Index associations with TwinsUK microbiota and health

To assess the ability of the HMI, MDI, and Shannon diversity to represent wide-scale microbiota composition in TwinsUK, I carried out PERMANOVA analysis using the adonis function within the vegan package in R (Oksanen et al. 2016). This was used to determine the association between the indices and both the weighted and unweighted UniFrac beta diversity measures between all samples. This was carried out using 1000 permutations to calculate pseudo p-values and estimate the significance of the associations.

To assess index associations with host health, the HMI, MDI and Shannon diversity were compared by their association with the frailty and number of disorders of individuals. Frailty was quantified using the Rockwood frailty index as described in Chapter 3. Both frailty and number of disorders followed a gamma distribution and were root normalised. The association between indices and health measures and Shannon diversity was assessed using linear regression.

6.2.9 Optimising the HMI

To optimise the HMI I used the glmnet package in R to perform Lasso variable selection, modelling the ability of all the marker families (and all the families within marker classes) to predict the full HMI. This identified a subset of families that significantly predicted the HMI. Using only this subset I generated an optimized HMI and compared it to the full index using Pearson correlation. The association of the optimised HMI with the root

normalised frailty index and number of disorders was assessed using linear regression as for the full index.

6.2.10 Replication datasets

Three publicly available 16S rRNA gene sequencing data sets were obtained for replication of the HMI. These were from faecal swabs from the American Gut cohort, a healthy population based cohort containing a participants from around the world but largely from the United States (*The American Gut Project*); the data from a study of paediatric IBD patients that utilised ileal bowel biopsies, chosen as it was used to define the MDI (Gevers et al. 2014); and data from a bowel cancer study that was selected to include a novel disease (Baxter et al. 2016).

American Gut

The American gut cohort data was obtained from the public FTP (ftp://ftp.microbio.me/AmericanGut) in the form of a pre-made OTU table and pre-calculated weighted and unweighted UniFrac distances (*The American Gut Project*). The OTU table was subset to samples that were listed within the faecal UniFrac table (which contained 6108 samples at the time of access) and with at least one data point in the three diseases considered (IBS, IBD, and diabetes), producing a final set of 5494 samples. This table was rarefied to 10,000 reads per sample prior to collapsing to family counts. The HMI and MDI were calculated as for TwinsUK. Shannon diversity was assessed using the alpha_diversity.py script within QIIME. Metadata relating to well defined diseases was selected from the large public mapping file. The PCoA axes for beta diversity plots were generated from UniFrac tables as for TwinsUK.

Gevers *et al.*

16S rRNA gene sequencing data from the Gevers *et al.* paper was downloaded from the QIITA repository (Study ID:1998) (Gevers et al. 2014). This was in the form of a closed reference OTU table created against the Greengenes reference and contained 1359 samples. This was subset discarding faecal samples to only retain biopsies and rarefied to a depth of 10,000 reads per sample producing a final set of 836 biopsy samples used for beta diversity analyses. This OTU table was collapsed to family counts and the indices generated as before. The appropriate Greengenes (v13_8) phylogeny was used to generate UniFrac distances, which were in turn used to generate beta diversity axes. Shannon diversity was assessed using the alpha_diversity.py script in QIIME. Disease metadata was also obtained from QIITA.

Baxter *et al.*

Data for the 490 samples in the Baxter *et al.* study were obtained from the European Nucleotide Archive (Study ID:PRJNA290926) (Baxter et al. 2016). Reads were collapsed to OTUs using close reference clustering using UCLUST within QIIME against the Greengenes reference database and rarefied to 10,000 reads per sample. Where multiple sequencing runs were available for a sample, only one was retained leading to the exclusion of 7 samples whose runs contained less than 10,000 observations. Family counts were generated from the OTU table allowing calculation of the HMI and MDI. Shannon diversity and UniFrac beta diversity distances were calculated as for the Gevers *et al.* data, again generating PCoA axes for plotting.

6.2.11 Index associations in replication data

Associations between the HMI, MDI and Shannon diversity index and disease states in replication datasets was assessed using one-tailed Mann-Whitney U-tests to determine if the healthier individuals were significantly higher scoring. In the American Gut dataset cases and controls in the comparisons were selected from the disease columns where 'I do not have this condition' was considered control and 'Diagnosed by a medical professional (doctor, physician assistant)' was considered a case, other samples were ignored on a per disease basis. In the case of the Gevers *et al.* data only the 482 ileal biopsy samples were considered in the disease comparisons. This limited the comparisons to one sample per person, maximised sample size (as it was the most sampled site), and was the tissue used in the original publication to define the MDI (Gevers et al. 2014). In the Baxter *et al.* study, accompanying metadata was used to classify disease status with samples listed as 'normal' and 'high risk normal' being considered healthy controls, and those with 'adenoma', 'advanced adenoma', and 'cancer' being considered disease cases. Plots of the first two axes of both the weighted and unweighted beta diversity PCoAs were coloured by the HMI, MDI, and Shannon indices for all data sets. PERMANOVA was used to assess the association between UniFrac measures and the HMI, MDI and Shannon diversity for all datasets as for TwinsUK.

6.2.12 Host influences on the HMI

Heritability

Heritability of the HMI was established using the mets package in R (Scheike, Holst, and Hjelmborg 2013). Heritability of the index was modelled fitting E, CE, AE, and ACE models; all adjusting for age, BMI, gender, and technical confounders including collection method

and sequencing depth. The best fitting model was that which had the lowest Akaike information criterion.

Discordant Twins

Discordant twin pairs for the root transformed frailty index and number of disorders were identified - where one twin scored higher than the mean of all individuals, the other scored lower, and the difference between the twins was at least twice the standard deviation across all samples. A one-tailed Wilcoxon signed-rank test was then used to establish if the healthier twins had significantly higher HMI scores.

Longitudinal Samples

Samples at two time points were available for a subset of twins. Individuals whose samples were at least 3 years apart were selected and HMI scores were generated for both samples as described previously. Pearson correlation was used to compare earlier and later time points within individuals, and an overall reduction in HMI score over time was investigated using a one-tailed Wilcoxon signed-rank test.

Diet

Weekly dietary nutrient intakes were estimated in the cohort from food frequency questionnaires, where twins provide a self-reported estimation of how often, on average, they consumed a specified amount of food over a year (Teucher et al. 2007). The association between dietary components and the HMI was determined using mixed effects models with weekly nutrient intake as the response, and age and zygosity as fixed covariates, and gender and technical variables (sequencing run, individual who loaded extraction plate, individual who extracted sample, and collection method) as random effects. Multiple testing was accounted for using FDR correction of p-values.

6.2.13 Metagenomic and metabolomic enrichment analysis

Metagenomic Pathways

Shotgun sequencing and metagenomic analysis had been carried out from the same stool samples as the 16S rRNA gene sequencing for a subset of 218 individuals by collaborators at the BGI. The DNA extraction, sequencing, and analysis of this metagenomic data is described in Section 2.3.1 (Xie et al. 2016). KEGG gene ontology annotations generated by the previous analysis were converted to relative abundances and log transformed, discarding KEGG Ortholog (KO) entries found in less than 25% of the original metagenomics

samples. Linear regression was then used to gauge associations between the HMI and log transformed relative abundances of each KO entry, including age and BMI as covariates. The KEGG REST API was used to obtain member pathways for each of the KO entries. Gene set enrichment analysis was carried out based on the KOs' beta coefficients with the HMI to identify enriched KEGG pathways, using the piano package in R (Väremo, Nielsen, and Nookaew 2013). Similar metagenomic pathway enrichment analyses were also carried out for the frailty index and number of disorders an individual reported (both measures root transformed). Enrichment was considered significant where multiple-testing corrected $p<0.05$.

Faecal Metabolite Groups

Faecal metabolomics profiling (Section 2.3.2) was available for a subset of 685 faecal samples with gut microbiota profiles. Similarly, to the metagenomic analysis, linear regression was used to identify associations between individual metabolites and the HMI. Enrichment for super groups of metabolites was then investigated as for metagenomic pathways using gene set enrichment analysis as before. Faecal metabolite enrichment analysis was carried out against the HMI, the frailty index (n=527) and number of disorders individuals had reported.

6.3 Results

6.3.1 Common disorders and medications in TwinsUK

Self-reported status for 38 common health disorders was assessed for 2737 individuals with gut microbiota profiles within the TwinsUK cohort (Table 6.1 and Figure 6.3 A & B). These included GI, autoimmune, and neurological disorders amongst others. Within the study population 96% reported affliction with at least one disorder (median 4, range 0-17), reflecting the elderly nature of the cohort. Correlation between disorders was generally low (Figure 6.3C) with some expected exceptions. For example, allergy correlated with eczema and asthma, consistent with the concept of atopy; and there was clustering of a group of disorders related to metabolic syndrome including hypercholesterolemia, ischaemic heart disease, hypertension, and T2D.

Self-reported medication use was available for 1724 individuals for 51 common drug-classes and antibiotic use within the month prior to sample collection was recorded for 2030 individuals (Table 6.2 and Figure 6.4 A & B). Where data was available, individuals reported a median use of 2 medications (range 0-21). Similarly to diseases, the overall correlations across medications were modest, with some predictable exceptions, such as calcium and

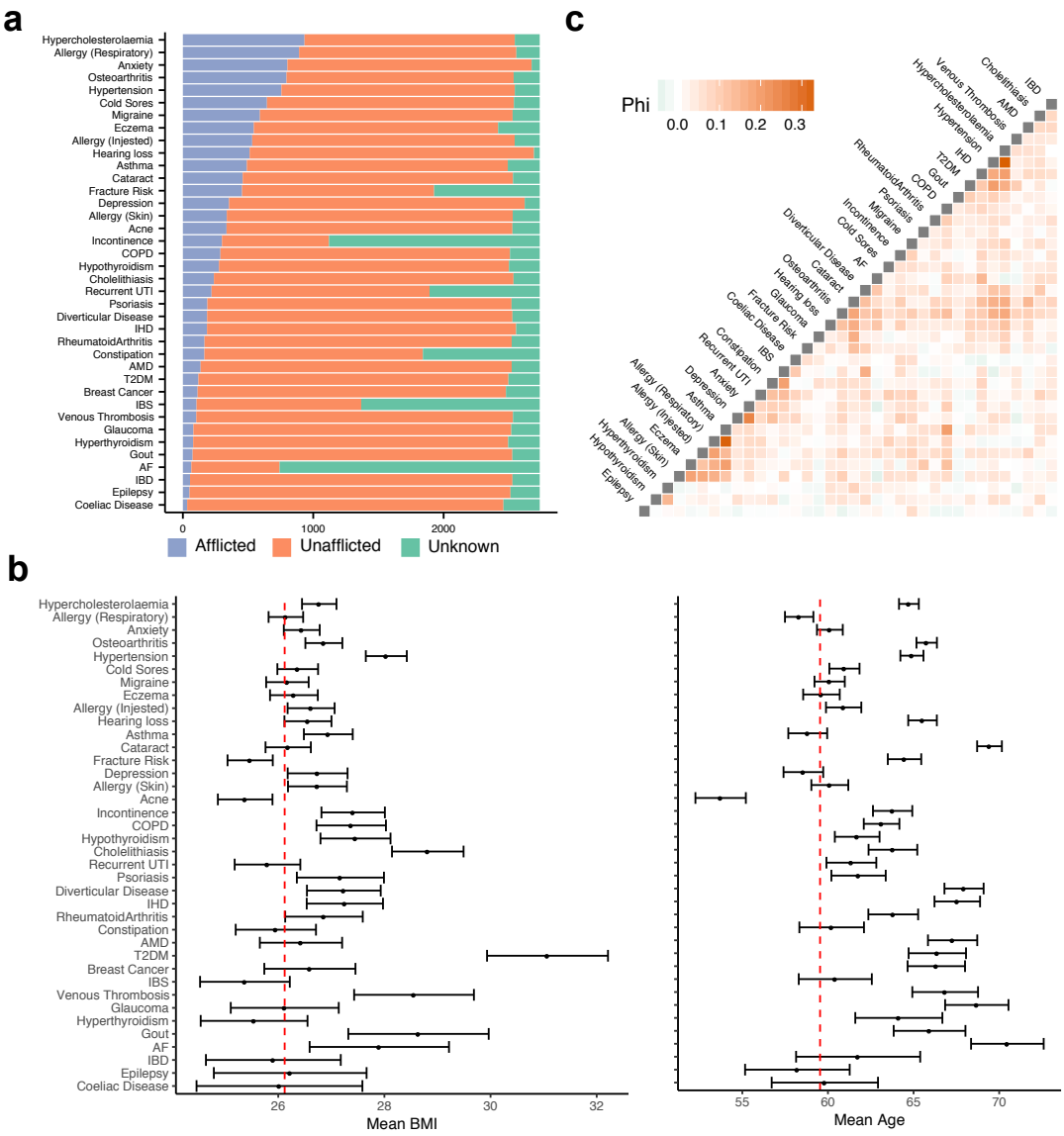


Figure 6.3: Summary of the common disorders considered in this analysis. A) The numbers of afflicted and unafflicted individuals for each disorder. Unknown represents missing data. B) The mean and 95% confidence intervals for BMI and age within the cases of each disorder. Overall means shown in red. C) Heatmap showing the correlation between disorder occurrence. The Phi coefficient is equivalent to Pearson's coefficient for binary variables.

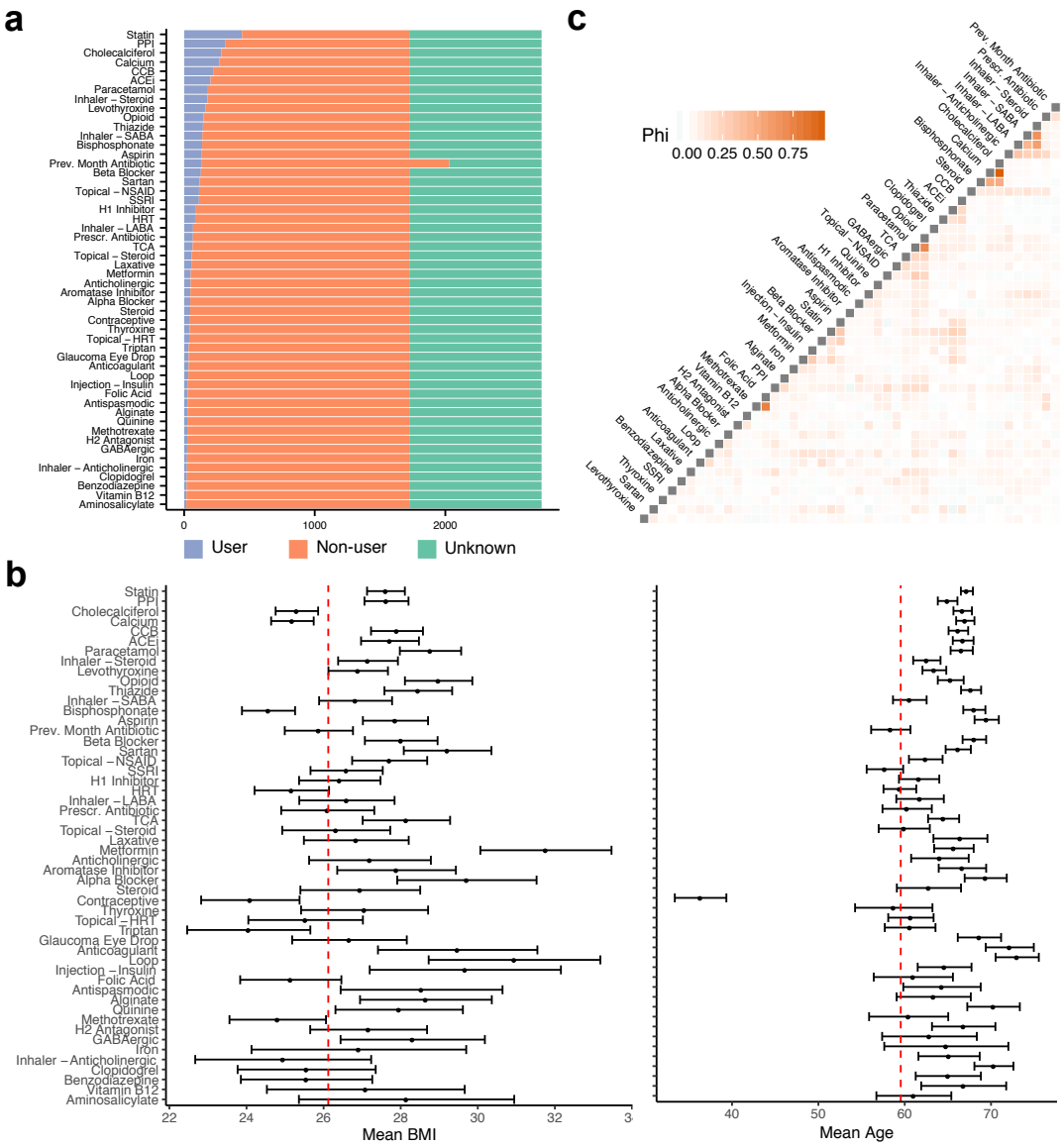


Figure 6.4: Summary of the common medications used in the cohort and considered in this analysis. A) The numbers of users and non-users for each medication. Unknown represents missing data. B) The mean and 95% confidence intervals for BMI and age within the users of each medication. Overall means shown in red. C) Heatmap showing the correlation between use of different medications. The Phi coefficient is equivalent to Pearson's coefficient for binary variables.

cholecalciferol that are commonly used in combination to treat conditions such as osteoporosis (Φ=0.97) (Figure 6.4C).

6.3.2 Disorder and drug associations with marker traits in the gut microbiome

Identifying marker traits

Gut microbiota profiles were generated from 16S rRNA gene sequencing using SUMACLUST as described in Chapter 5. To reduce the statistical burden of multiple testing associated with assessing numerous traits in one study, I used a heuristic approach to select a subset of

marker features for microbiota analyses (Figure 6.2). This approach identified 68 marker traits that represented wider microbiota composition. The final set included Shannon and phylogenetic diversity measures of alpha diversity, most of the beta diversity principal co-ordinate axes (from the first 6 axes of PCoA using both weighted and unweighted UniFrac distances), and a range of taxonomic abundance summaries from both the family and class level (Table D.1).

Disorder associations

Association analyses between disorder status and microbiome markers, correcting for age and body mass index (BMI), identified nominally significant associations ($p<0.05$) for almost all disorders (Figure 6.5A and Table D.2). Ingested allergy status had the most associations, followed by inflammatory bowel disease (IBD) and type-2 diabetes (T2D). After false discovery (FDR) correction (5%), 17 disorders retained at least one significant association. As the number of cases determines the power to detect associations, I compared the number of microbiome associations relative to the number of disorder cases (Figure 6.6A). T2D, IBD, constipation, recurrent cystitis, and coeliac disease had relatively more associations than more common disorders, suggesting a stronger association with the gut microbiome.

From the microbiota marker traits, principal co-ordinate axes derived from beta diversity measures had the most associations with disorders (Figure 6.5B). Alpha diversity had exclusively negative associations with several disorders, in concordance with the reduced diversity observed with frailty in Chapter 3 and multiple reports of reduced gut microbiome diversity with chronic conditions such as IBD and T2D (Gevers et al. 2014; Qin et al. 2012).

Obesity was omitted in these analyses as I wanted to determine effects independent of BMI, which has been comprehensively studied in relation to the gut microbiota (Goodrich et al. 2014b; Turnbaugh et al. 2006). However, for comparison, repeating the analysis without adjustment for BMI, obesity (BMI >30) had the most (n=27) FDR significant associations of all the conditions (Figure D.2).

Drug associations

There were nominally significant associations with microbial markers for 51 of the 52 drugs, and 19 had at least one association after FDR adjustment (Figure 6.5C and Table D.3). Comparing the number of observed associations with the number of medication users (Figure 6.6B), antibiotics, anticholinergic inhalers, SSRIs, paracetamol, and opioids had proportionally more associations. As for diseases, most of the microbiota associations with drugs were with beta diversity principal components and alpha diversity measures (Figure 6.5D).

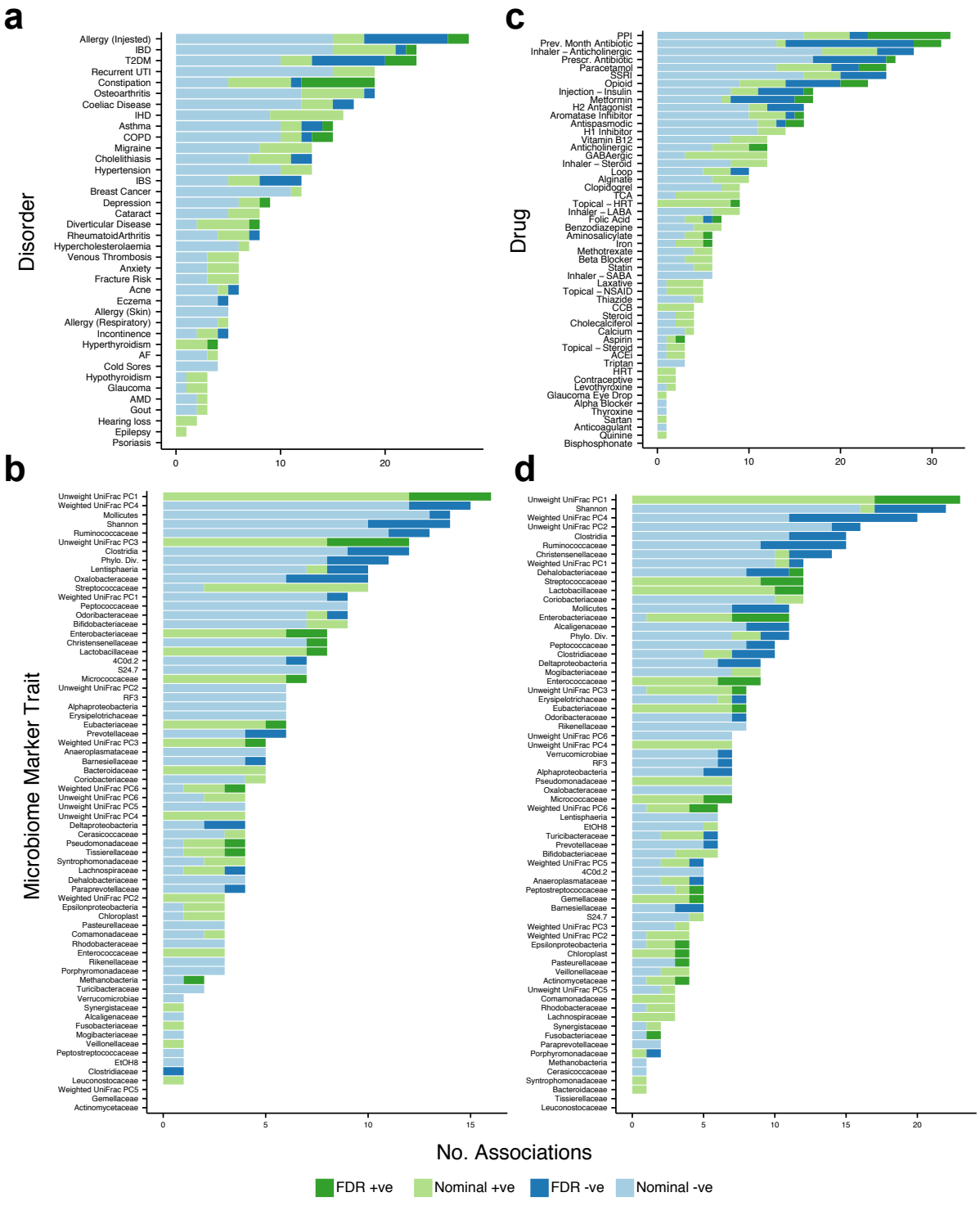


Figure 6.5: Associations between diseases and drugs and the microbiota markers. The number of nominally and FDR significant associations observed with the microbiota marker traits and each disorder (A) and use of each drug (C). B and D show the number of associations with each microbiota marker trait across all the disorders and drugs considered respectively.

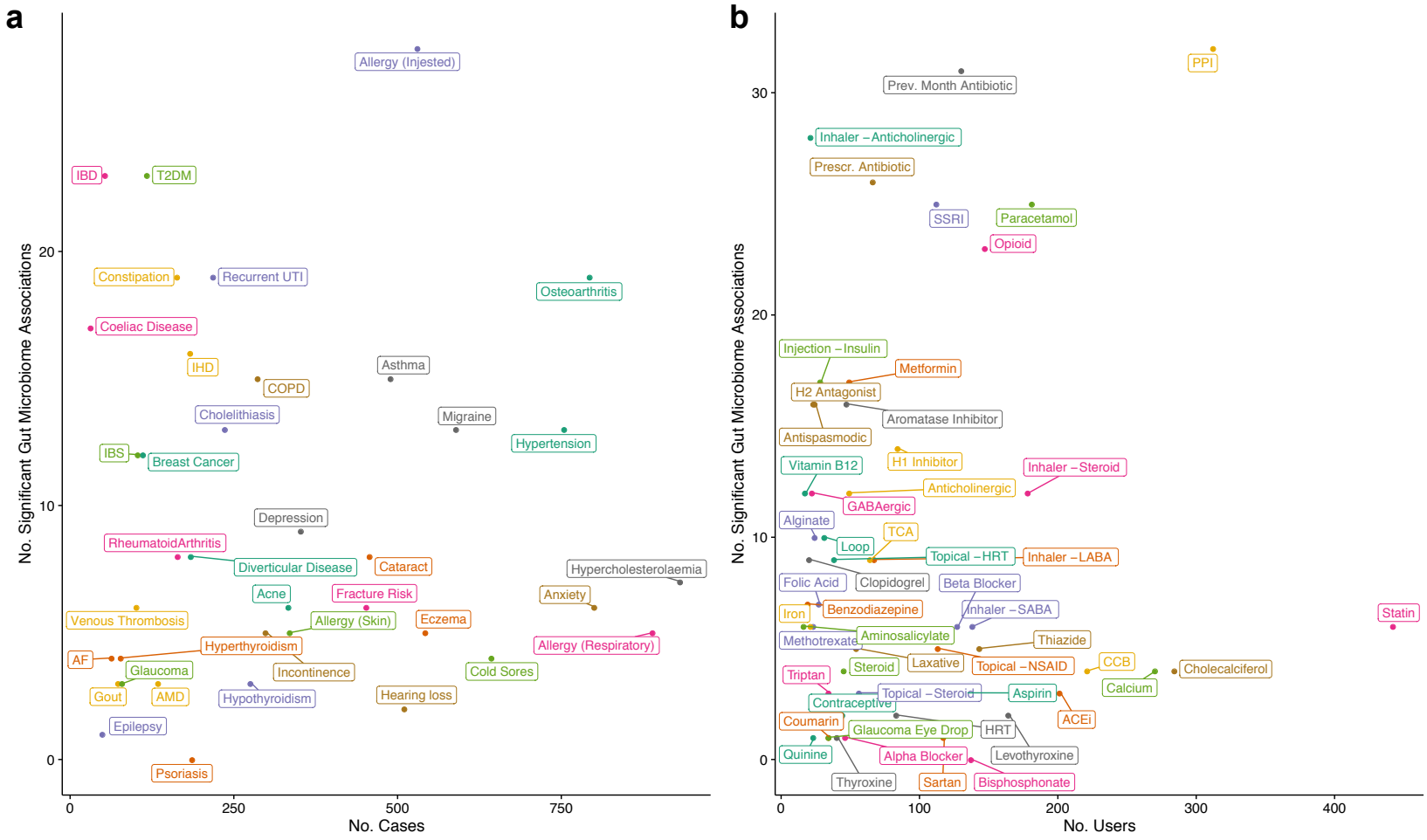


Figure 6.6: Number of cases compared to the number of associations observed. The number of nominally significant associations observed across all microbiota marker traits plotted against the number of cases/users for disorders (A) and drug use (B).

6.3.3 Delineating disorder-drug effects

An often-overlooked challenge in observational studies is the delineation of association specificity to either a disorder or its treatment. To attempt to address this, I inferred a Bayesian network of the statistical dependencies between each disorder, drug, and microbial trait (Figure 6.7). In the subset of 1724 individuals with complete data, there were 272 conditional dependencies (Table D.4), 38% of them between microbial features, 8% between disorders, 20% between drugs, and 22% between disorders and drugs. There were 15 instances where a microbial trait was statistically dependent on disorder status, independently of medications and other disorders. Five of these associations replicated nominally significant associations in the previous pairwise association analyses and two, Constipation-Methanobacteria and Diverticular Disease-Pseudomonadaceae, were previously significant after FDR 5% adjustment. There were 17 instances of microbial features dependent on drug use, five were previously observed as nominally significant associations and three as FDR significant: PPI-Streptococacceae, Metformin-Clostridiaceae, and Aromatase Inhibitors-Enterococcaceae.

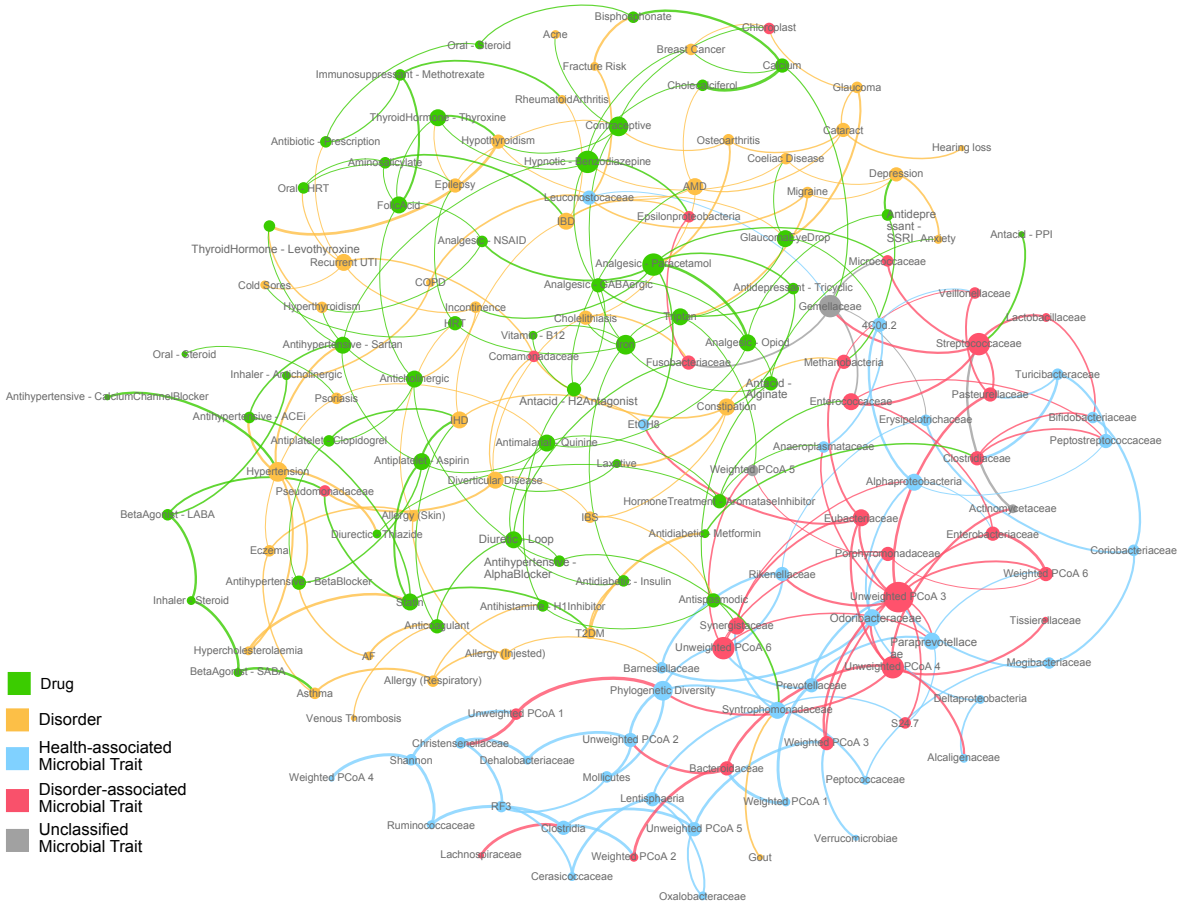


Figure 6.7: Bayesian network inferred from disorder, drug-use, and gut microbiome data. Nodes represent either disorders, drugs, or microbiome traits. Edges represent statistical dependence between the nodes, and are coloured by the source node. Edge weights are relative to the strength of the dependency. The health and disorder definitions are based on the clustering in Figure 6.8

6.3.4 Defining a health-associated gut microbiome

To capitalise on the comparability of the association results, I compared the direction of the significant associations between them. Unsupervised clustering of the 68 microbial markers by their coefficients of association with diseases identified two clusters, in which the marker traits were generally either positively or negatively significantly associated with more than one disorder (Figure 6.8A). Bootstrap evaluation of clustering confirmed the significance of the clusters ($p<0.05$), which I use to classify the markers into two groups: disorder-associated (cluster one: higher abundance associated with disorders) and health-associated (cluster two: lower abundance associated with disorders). Beta diversity principal co-ordinate axes were distributed between the two groups, while alpha diversity measures were all health-associated. Within taxa, classes that also had markers at the family level displayed consistent groupings (Figure 6.8B). The exception was Clostridia, a health-associated class that encompassed marker families associated with both health and disorders - in keeping with common assent that Clostridia is a polyphyletic taxon.

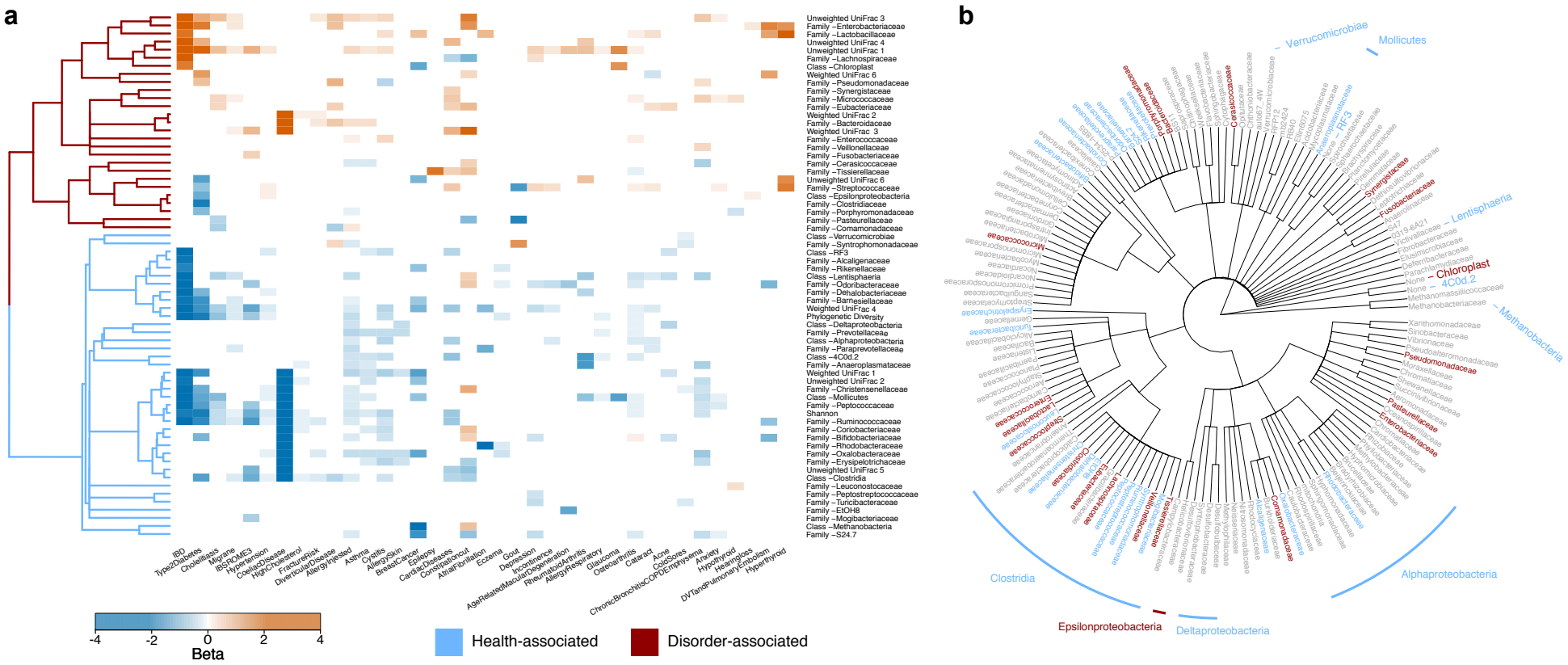


Figure 6.8: Defining a universal health-associated gut microbiome. a) Hierarchical clustering of microbiota marker traits by beta coefficients from nominally significant disorder associations. Beta of non-significant associations were assigned 0 and hence coloured white. Complete hierarchical clustering of cosine distances between both microbiome traits and disorders was used to generate the heat map shown. Bootstrap clustering of microbiome traits identified two significant clusters highlighted in the left dendrogram. These are used to define the health and disorder associated traits. b) Taxonomic plot of all families with complete taxonomic assignment observed in the cohort. Marker families used to define the HMI are coloured by their health/disorder associated grouping. Marker classes are also highlighted around the periphery. Tree structure is based on taxonomic not phylogenetic relationships.

6.3.5 The health-associated microbiome index (HMI)

Based on the groupings of microbial families and classes in Figure 6.8, I defined a health-associated gut microbiome as one with a higher abundance of health-associated taxa relative to disorder-associated; and quantified this in a single score following the approach of Gevers et al., who defined the Microbial Dysbiosis index (MDI) to compare bacteria differentially associated with Crohn's disease (Gevers et al. 2014). This score is the log10 of the ratio of health-associated to disorder-associated taxa, which I have termed the Health-associated Microbiome Index (HMI). Of the ten taxa comprising the MDI, six were also included in the HMI and all had consistent direction of association.

HMI associations with beta diversity and host health

I used PERMANOVA to compare the ability of the HMI, MDI and the Shannon diversity index to represent wide-scale microbiota composition within TwinsUK (Table 6.3 & Figure 6.9). All three significantly associated with compositional distances as measured by both the weighted and unweighed UniFrac distance. However, the HMI represented a larger proportion of the variance than both the MDI and Shannon diversity in the weighted measure. The variance explained by the HMI was on par with Shannon diversity when using the unweighted UniFrac measure. Comparing the indices association with health measures, the HMI also associated more strongly than the MDI with the number of disorders an individual reported and with their frailty level (Figure 6.10).

Table 6.3: PERMANOVA results for index associations with UniFrac measures in TwinsUK. All associations were significant (p=0.0009, minimum in 1000 permutations).

Index	UniFrac Measure	R^2
HMI	Unweigthed	0.02027
HMI	Weighted	0.14276
MDI	Unweigthed	0.00354
MDI	Weighted	0.02475
Shannon	Unweigthed	0.02622
Shannon	Weighted	0.06035

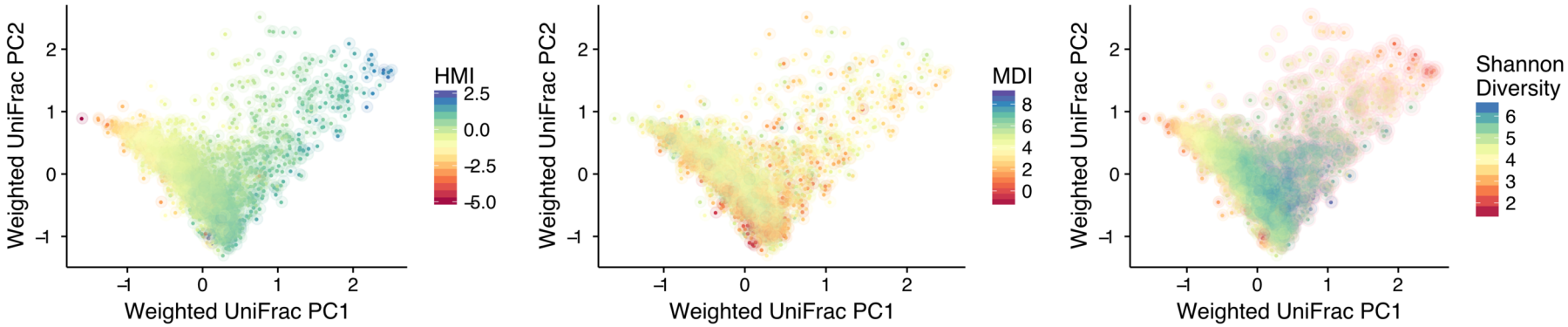


Figure 6.9: Beta diversity plots showing the first two axes from PCoA analyses using the weighted UniFrac metric across the 2737 samples in TwinsUK. Plots are coloured based on the HMI, MDI, or Shannon diversity. On the Shannon plot there is additional colouring by *Prevotella* abundance (red circles), highlighting the out group of individuals with low diversity whose microbiota are dominated by *Prevotella*.

HMI and alpha diversity

The HMI was also positively correlated with alpha diversity (Figure 6.10). However, a subset of samples had a high HMI but low diversity and were characterised by a high abundance of taxa from the genus *Prevotella* (Figure 6.9 and 6.10) . Within TwinsUK, *Prevotella* rich individuals had a significantly lower frailty index and number of disorders than other members of the cohort ($p<0.001$ in Mann-Whitney U tests)(Figure 6.11). The *Prevotella* dominant microbiome therefore represents a healthy but low diversity state countering the typical trend across other samples in this dataset. The HMI more accurately represented health status than Shannon diversity within the *Prevotella* rich subset.

An optimised HMI

Lasso variable selection was used to identify a minimal set of microbiota marker families required to define the HMI. This found 17 health and 8 disorder-associated families that independently contribute to the index (Table 6.4). An optimised index created only using this subset was highly correlated with the full HMI ($r=0.99$). The families dropped from the full HMI had significantly lower abundance in the cohort (Mann-Whitney U test, $p<0.001$), and hence little influence on the index. There was no significant difference between the abundances of the health-associated and disorder associated families in the optimised HMI (Mann-Whitney U test, $p=0.34$), suggesting the HMI is not simply quantifying the ratio of rare to abundant taxa. The optimised HMI retained significant associations in linear regressions with number of disorders ($p<0.001$) and frailty ($p<0.001$) of an individual, as expected given the high correlation with the full index. The families within this optimised index are therefore sufficient to recapitulate the HMI.

6.3.6 The robustness of the HMI

The HMI was strongly associated with microbiota composition and host health within TwinsUK. However, this might be expected given it was defined within the cohort and could be the result of over fitting the data. To determine the robustness of the HMI I calculated it within three independent disease datasets and again contrasted its associations with the MDI and Shannon indices. I examined the indices associations with IBD (cases=157, controls=4837), IBS (cases=440, controls=2475), and diabetes (cases=69, controls=5142) from a wider set of 5494 faecal samples from the American Gut cohort (*The American Gut Project*); colon cancer (cases=313, controls=170) in 483 faecal samples from a study by Baxter *et al.* (Baxter et al. 2016); and IBD status in 482 ileal biopsy samples (cases=329, controls=153) from a wider set of 836 biopsies from several sites along the GI tract as described by Gevers *et al.* (Gevers et al. 2014).

Table 6.4: The familial marker traits used to define the optimised HMI. Lasso variable selection was used to identify the most significant predictors of the full HMI from all the family level markers and all the families within the class level markers. Shown are the features Greengenes taxonomic assignment and their health/disorder classification.

Association Group	Greengenes Family Assignment
Health	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__S24-7
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Barnesiellaceae]
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Odoribacteraceae]
Health	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Paraprevotellaceae]
Health	k__Bacteria;p__Firmicutes;c__Clostridia;Other;Other
Health	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;Other
Health	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__
Health	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Christensenellaceae
Health	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptococcaceae
Health	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae
Health	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae
Health	k__Bacteria;p__Lentisphaerae;c__[Lentisphaeria];o__Victivallales;f__Victivallaceae
Health	k__Bacteria;p__Tenericutes;c__Mollicutes;o__RF39;f__
Health	k__Bacteria;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Verrucomicrobiaceae
Disorder	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae
Disorder	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae
Disorder	k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__
Disorder	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae
Disorder	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae
Disorder	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae
Disorder	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae
Disorder	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae

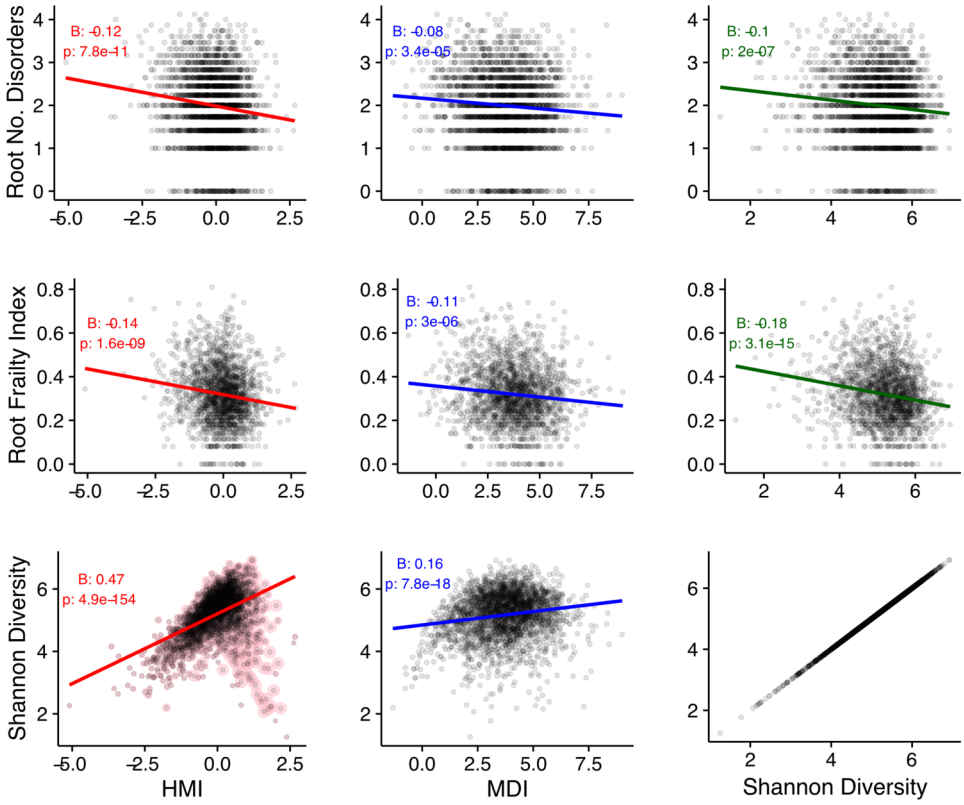


Figure 6.10: Plots showing the association between the HMI, MDI, and Shannon indices, with the number of disorders and frailty of an individual and the Shannon diversity of their gut microbiota. Shown are the beta coefficients and p-values for their associations as determined using linear regression. In the HMI and Shannon plot, the *Prevotella* rich group are highlighted as for Figure 6.9.

Robust HMI-health associations

The HMI was consistently significantly higher in healthy individuals in all cases except the colonic cancer study (Figure 6.12A). Shannon diversity was significantly higher in healthy individuals in the IBD and IBS comparisons, but not the diabetes and colonic cancer comparisons. The MDI was higher in healthy individuals in all cases except the colonic cancer study and the American Gut IBD comparison. The HMI did not have the largest difference between means in the index-disease comparisons. For instance, the MDI was most disparate between cases and controls in the Gevers *et al.* ileal IBD dataset, within which it was defined, and the Shannon index had the largest difference in the American Gut IBD comparison. However, the HMI was the most consistent index, producing a significant difference in the expected direction across all the comparisons except in the Baxter *et al.* data.

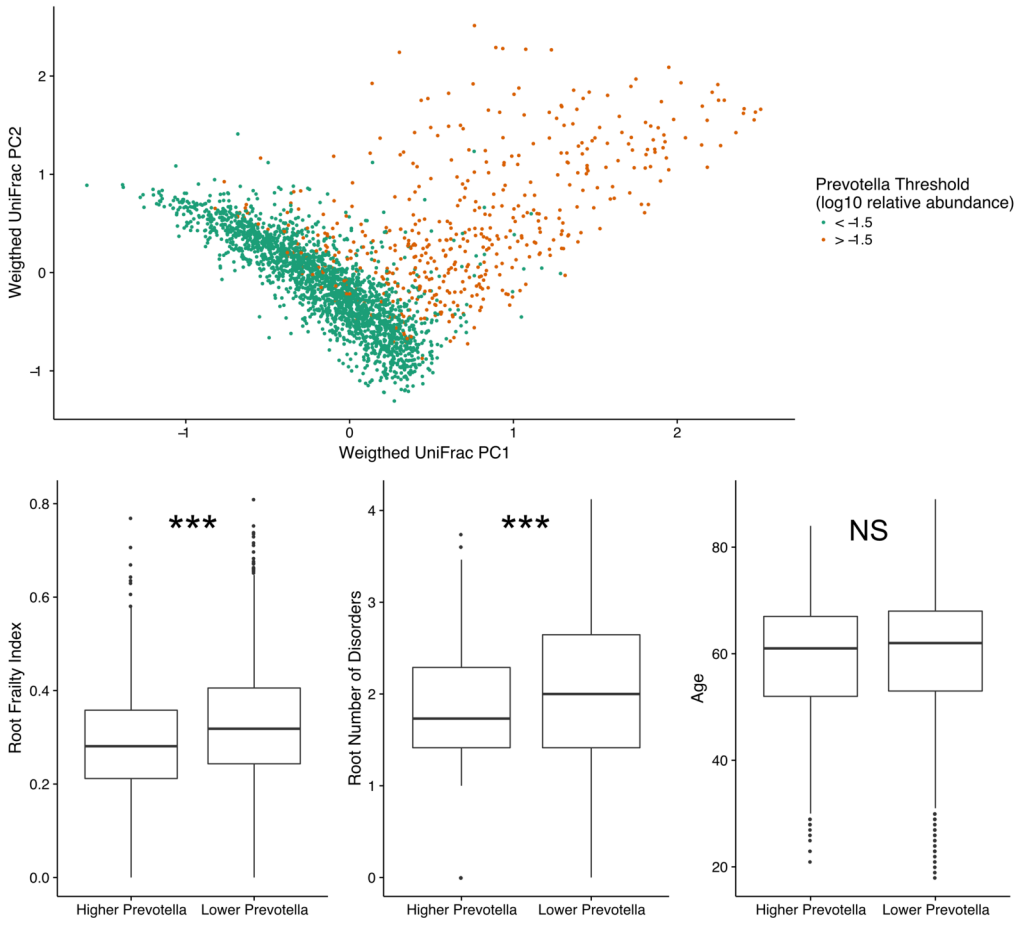


Figure 6.11: TwinsUK members with a *Prevotella* rich microbiome have low diversity but are significantly healthier. The top plot shows the cut-off used to divide samples as Prevotella dominant, it was selected as it encompassed the main out-grouping on the beta diversity plot. Below, Mann-Whitney U tests comparing the frailty, number of disorders and age differences between the two groups (*** = p<0.001).

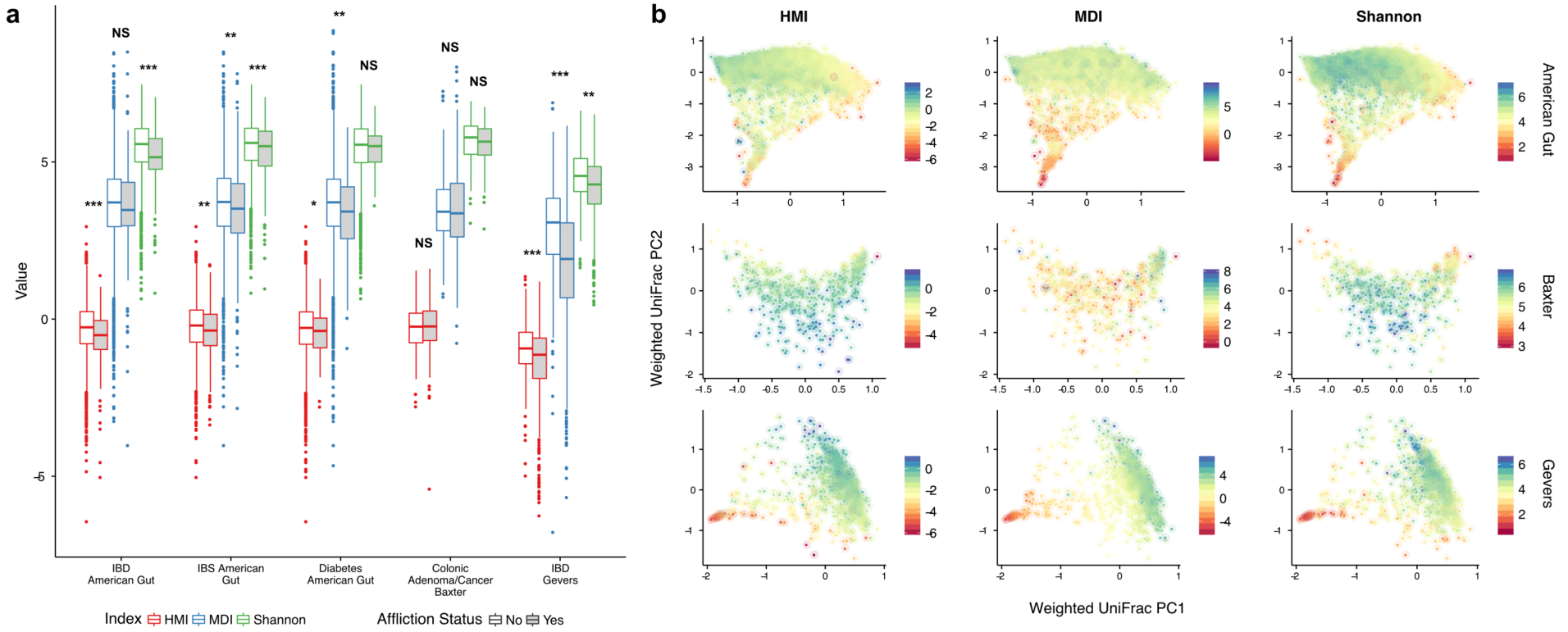


Figure 6.12: Health and wide-scale microbiota composition associations of the HMI replicate in external datasets. A) Comparison of the HMI, MDI and Shannon index between the health and diseased state in 5 disorders from 3 publically available data sets. American Gut - IBD (157, 4837), IBS (440,2475), Diabetes (69,5142); Gevers *et al.* (329,153); Baxter *et al.* (313,170) (no. cases, no. controls). ***=p<0.001, **=p<0.01, *=p<0.05 in one-tailed Mann-Whitney U tests. B) PCoA plots from weighted UniFrac beta diversity measures from three public data sets colored by the HMI, MDI, and Shannon index demonstrates the compositional representation of the HMI across all three. American Gut (n=5494), Gevers *et al.* (n=836), and Baxter *et al.* (n=483).

Consistent HMI associations with beta diversity

Principal coordinate plots of the beta diversity within these studies (Figure 6.12B), showed that the MDI could summarise overall microbiome composition well in the Gevers *et al.* and American Gut data, but less so in the TwinsUK (Figure 6.9) and Baxter sets; whilst the HMI matched compositional patterns well in all four studies. These results were reflected in PERMANOVA analyses of the HMI, MDI, and Shannon index associations with UniFrac distances in these datasets (Table 6.5). In all but the Gevers *et al.* dataset, in which the MDI was defined, the HMI explained more variance than the MDI in both the weighted and unweighted UniFrac distances. The HMI explained more variance than Shannon diversity in all the weighted comparisons except within the Baxter *et al.* dataset. Although in all comparisons, the HMI and Shannon diversity indices explained a comparable proportion of the UniFrac variance. These results show that the HMI is a robust index that can be used to represent wide-scale microbiota composition with a directional association to host-health across multiple populations.

Table 6.5: PERMANOVA results for index associations with UniFrac measures in replication datasets. All associations were significant (p=0.0009, minimum in 1000 permutations).

Index	Dataset	UniFrac Measure	R^2
HMI	American Gut	Unweigthed	0.031
HMI	American Gut	Weighted	0.132
MDI	American Gut	Unweigthed	0.013
MDI	American Gut	Weighted	0.048
Shannon	American Gut	Unweigthed	0.045
Shannon	American Gut	Weighted	0.113
HMI	Gevers *et al.*	Unweigthed	0.055
HMI	Gevers *et al.*	Weighted	0.219
MDI	Gevers *et al.*	Unweigthed	0.054
MDI	Gevers *et al.*	Weighted	0.387
Shannon	Gevers *et al.*	Unweigthed	0.065
Shannon	Gevers *et al.*	Weighted	0.202
HMI	Baxter *et al.*	Unweigthed	0.037
HMI	Baxter *et al.*	Weighted	0.064
MDI	Baxter *et al.*	Unweigthed	0.010
MDI	Baxter *et al.*	Weighted	0.059
Shannon	Baxter *et al.*	Unweigthed	0.055
Shannon	Baxter *et al.*	Weighted	0.075

6.3.7 Host influences on the health-associated gut microbiome

To demonstrate the utility of the HMI, I used it to investigate host factors that determine a proclivity to a more health or disorder-associated gut microbiome.

Gender

There was no significant difference in the HMI by gender in TwinsUK (Mann-Whitney U test, p=0.15), Baxter *et al.* (p=0.28), or Gevers *et al.* (p=0.08) data. In the larger American Gut cohort, we observed a small but significantly higher HMI score in males relative to females (p=0.001).

Genetics and environmental influences on the HMI

Using the twin structure of the TwinsUK cohort, I estimated the genetic influence (heritability) of the index. The best fitting model (AE) attributed 34% (95% CI=29-39%) of HMI variance to additive genetic effects, which is a high heritability relative to other microbiome measures (Goodrich et al. 2014b). Extending the twin analysis, I compared the HMI between twin pairs discordant for both frailty (n=47) and number of disorders (n=53). Less healthy twins scored significantly lower in both cases (Figure 6.13A), suggesting an association with health independent of genetic effects. This is consistent with evidence that the gut microbiome is determined by both genetic and environmental factors (Consortium 2012; Goodrich et al. 2014b). Longitudinal stool samples were collected at two time points (either 3 or 4 years apart) for 349 individuals (Figure 6.13B). The earlier HMI score was significantly correlated with the later (r=0.35, p<0.001) and later time points were significantly lower scoring than their earlier counterparts (p<0.001), which might be expected assuming a decline in health with age.

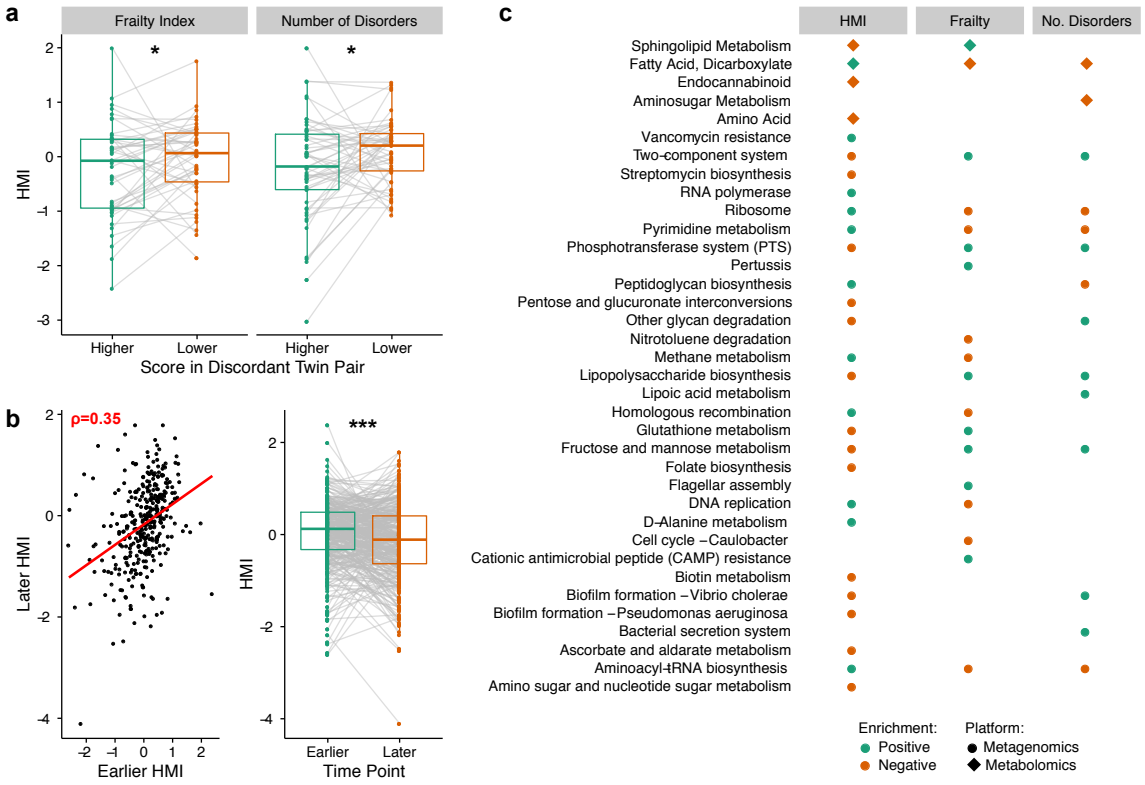


Figure 6.13: Using the HMI to identify host influences on, and functional properties of, the health-associated microbiome. a) Comparison of HMI scores in twin pairs discordant for frailty and number of disorders. *=p<0.05 in one-tailed Wilcox signed rank tests. b) Left, correlation between HMI scores of two samples from the same individuals at least 3 years apart. Right, HMI scores of later samples are significantly lower than in earlier samples. ***= p<0.001 in one-tailed Wilcox signed-rank test. c) Significant results (FDR adjusted p<0.05) in enrichment analysis for both metagenomic pathways and faecal metabolite groups with the HMI score, frailty, and number of disorders. A positive enrichment means the genes/metabolites are enriched with a higher HMI, a negative enrichment with a lower HMI. A higher HMI score is associated with health, whereas a higher frailty index score, or number of disorders indicates poor health.

Dietary associations with the HMI

Diet can influence both host health and the gut microbiome (Wu et al. 2011; Zhernakova et al. 2016). The association between the HMI and estimated weekly intake of macronutrients (fat, carbohydrates, protein, and fibre) was assessed in a subset of 2006 twins. Fibre intake was found to have a small, but significant positive association with the HMI (β=0.08, p=0.001), and remained the only significant association repeating the analysis at a higher resolution separating the macro-nutrients into subsets of specific sugars and fats (Table 6.6).

Table 6.6: Results of association analyses between dietary nutrient intakes and the HMI using two different nutrient summary levels.

Weekly macronutrient intake	β	p	FDR Adjusted p
Carbohydrate (g)	0.040	0.073	0.145
Fat (g)	-0.006	0.805	0.805
Fibre (g)	0.081	0.000	0.001
Protein (g)	0.020	0.376	0.501

Weekly nutrient intake	β	p	FDR Adjusted p
Alcohol (g)	0.023	0.304	0.470
Carbohydrate (g)	0.040	0.073	0.234
Cholesterol (mg)	-0.004	0.873	0.873
Fat (g)	-0.006	0.805	0.873
Fructose (g)	0.049	0.027	0.231
Glucose (g)	0.037	0.096	0.234
Lactose (g)	-0.005	0.839	0.873
Maltose (g)	0.042	0.070	0.234
Monounsaturated fats (g)	-0.007	0.762	0.873
Fibre (g)	0.081	0.000	0.005
Polyunsaturated fats (g)	0.038	0.091	0.234
Protein (g)	0.020	0.376	0.533
Saturated fat (g)	-0.032	0.152	0.322
Starch (g)	0.043	0.057	0.234
Sucrose (g)	-0.012	0.585	0.765
Total sugars (g)	0.027	0.227	0.429
Trans-unsaturated fatty acids (g)	-0.025	0.275	0.467

6.3.8 Functionality of the health-associated gut microbiome

Finally, I used the HMI to investigate the functional properties of the health-associated gut microbiome. In a sub-set of individuals with previously processed metagenomic data, gene set enrichment analysis was applied to KEGG pathway abundances to identify pathways enriched with the HMI (n=218), frailty index (n=203), and the number of disorders an individual had (n=213) (Table D.5). Enrichment analysis against these measures was also carried out for metabolite groups using faecal metabolite data from 685 individuals (n=527

with frailty data) (Table D.6). Several significant (FDR adjusted $p<0.05$) positive and negative enrichments were observed for both metagenomic pathways and faecal metabolite groups with the HMI (Figure 6.13C). In all significant observations, the directions of association were inverse between the HMI and the frailty and number of disorder metrics as would be expected. These associations are discussed further below.

6.4 Discussion

In this chapter I have presented a comprehensive study of gut microbiota associations with host health, encompassing 38 common disorders and 51 different medications. This was made possible by using an approach designed to address the limits of multiple testing through the identification of marker traits in the gut microbiota. The subsequent associations identified using these markers revealed novel microbiota associations and also enabled comparison between the results. These comparisons revealed microbiota with consistent directions of association across the disorders considered, allowing definition of a health-associated gut microbiota. I quantified this in the HMI, which proves to be a robust method to summarise the composition of the human gut microbiota with relevance to host health. The HMI provides a novel tool for future studies investigating the gut microbiome in relation to human health.

6.4.1 Associations with gut microbiota markers

The greedy approach used to identify microbiota marker traits in this chapter, provides a novel method to address the limitations associated with the high dimensionality of 16S rRNA gene sequencing profiles. It enabled association analyses across a wide range of disorders and medications, several of which have not previously been studied in relation to the gut microbiome. The association analyses between disorders and microbiota markers replicated several reported associations including a negative association between T2D and Clostridia (Larsen et al. 2010), a positive association between methanogens and constipation (Pimentel et al. 2012), and a lower abundance of Ruminococcaceae with IBS (Pozuelo et al. 2015). There were also several novel associations with disorders including urinary incontinence, acne, and osteoarthritis, which warrant further disease specific studies.

Within the medication analyses, most associations were with drugs known to influence the gut microbiome including antibiotics (Jakobsson et al. 2010), PPIs (Chapter 4), and metformin (Forslund et al. 2015). However, novel associations were observed including those with paracetamol, SSRIs and inhaled medications. This suggests that the major medication confounders for gut microbiome studies have been established, but highlights the importance of considering all treatments in gut microbiota studies as even non-oral medications could have an impact.

One of the major limitations of this study is the inability to delineate effects specific to a disorder, its treatment, or another similarly associated host factor. This was in part controlled for by adjusting for BMI and age. However, including further factors such as diet would subset the available datasets too much. To account for all the different diseases and treatments within regression models would be too complex given the number of different traits considered for both. The use of a Bayesian network here enabled some associations to be partitioned specific to either drugs or disorders and showed the potential of combinatorial multivariate approaches. However, the smaller subset with complete data provided much less power in this complex analysis. Probing the specificity of the associations to disease or treatment will require more targeted studies designed with populations selected to minimise the variation in other factors. However, these results provide guidance for where such efforts are best placed.

6.4.2 A health associated gut microbiota

Carrying out multiple disorder association analyses within the same cohort negated the issues of incomparability that would be met trying to compare results between studies. This enabled comparisons between the disorder associations and identification of generally health or disorder associated taxa in the gut microbiome. Within the health-associated bacterial families positive health outcomes have previously been associated with Christensenellaceae (Goodrich et al. 2014b), Rikenellaceae and Odoribacteraceae (Lim et al. 2016), Prevotella (Kovatcheva-Datchary et al. 2015), and Erysipelotrichaceae (Pozuelo et al. 2015). Within the disorder-associated families diseases have previously been associated with Bacteroides (Ng et al. 2013), and Veillonellaceae, Pasteurellaceae and Enterobacteriaceae (Gevers et al. 2014).

However, there are also contrasting associations. For example, a recent review by Lloyd-Price and colleagues collated a list of health associated families in the gut microbiota from a range of studies (Lloyd-Price, Abu-Ali, and Huttenhower 2016). This included Prevotellaceae, Ruminococcaceae, and Bifidobacteriaceae that were also health-associated in our study. However, it also included Bacteroidaceae, Clostridiaceae, Enterobacteriaceae, Eubacteriaceae, and Lactobacillaceae, all of which were disorder associated in our analyses.

Discrepancies in the direction of association of taxonomic groupings do not negate the validity of the health-associated definition proposed. There is significant taxonomic variation between healthy individuals (Human Microbiome Project Consortium, 2012), which suggests it is unlikely there are core taxa absolutely required for health. Furthermore, within species there can be pathogenic and non-pathogenic strains, and even the same bacteria could be commensal or pathogenic depending on environmental context (Ehrlich, Hiller, and Hu 2008). However, the health-associated definition is not intended to state that all members of the defining taxa are either health or disorder associated, but rather that higher

abundance of these broad level taxa reflect a gut microbiome more commonly associated with either end of the health spectrum at the population level. There are also analytical benefits to using higher taxonomic levels to define the health associated gut microbiome. It facilitates application across studies as broad level taxa are less influenced by the variability between them (as shown by the reduced variability between taxonomic summaries compared to OTUs in the clustering comparisons of Chapter 5).

6.4.3 The healthy microbiome index

I proposed the HMI to quantify the health associated microbiome. This was more strongly associated with measures of health and alpha diversity than the MDI in TwinsUK. It was also more widely applicable than the MDI in disease replications across the three public data sets and explained a greater proportion of the variance in beta diversity measures. This is expected as the MDI was not designed as a general measure of gut microbiome health and is specific to paediatric Crohn's disease (Gevers et al. 2014).

Microbiota diversity is also frequently used as a measure of health in the gut microbiota, on the basis that low diversity has been associated with a range of health deficits (Gevers et al. 2014; Qin et al. 2012). The HMI proved comparable to the Shannon diversity index, in that it represented a similar proportion of the variance of beta diversity measures. However, the individuals with a Prevotella dominant gut microbiome in TwinsUK were healthier but had low alpha diversity; whereas their HMI score was high - maintaining a consistent direction of association with health. Furthermore, the HMI has benefits over Shannon diversity in the taxonomic basis of its definition. Diversity measures are influenced by factors such as OTU clustering method that are less influential on higher level taxonomic abundances (Chapter 5) (Westcott and Schloss 2015). The taxonomic basis also enables comparable enumeration across different gut microbiota studies. The HMI is also specific to the human gut microbiome unlike ecological indices naive to taxonomy.

The HMI addresses typical OTU limitations of reproducibility, dimensionality, and clinical relevance; by providing a robust measure to quantify wide-scale gut microbiome composition within a single measure that has a directional association with host health. With further development it could be applied as an outcome in intervention studies aiming to improve the gut microbiota, as a diagnostic of health, or used to predict patient responses to treatments targeting the gut microbiome. I have also demonstrated its utility as a research tool by investigating the function of the health-associated microbiota using the HMI.

6.4.4 Functions of a health-associated gut microbiota

Using the HMI to carry out functional enrichment analyses with metagenomic and faecal metabolomic data, I was able to probe the functionality of a health-associated gut micro-

biota. I observed several significant enrichments for both metagenomic pathways and metabolic groups, some of which are discussed below.

One key mechanism by which the gut microbiome could influence a range of host health effects is through dysregulation of the immune system leading to low level GI and/or systemic inflammation (Lynch and Pedersen 2016; Thevaranjan et al. 2017). This could in turn modulate inherent polygenetic susceptibility to multiple diseases (Lynch and Pedersen 2016). The HMI enrichment analyses identified significant associations with several gut microbiome functions previously linked to GI inflammation.

Microbial glycan degradation genes were negatively associated with the HMI. Degradation of mucin glycans by gut microbiota in mice undergoing dietary fibre deprivation compromises GI barrier integrity promoting subsequent pathogen infection (Desai et al. 2016). Lipopolysaccharide production was negatively associated with the HMI. Lipopolysaccharides (LPS) are cell membrane components of gram negative bacteria that can infiltrate the GI epithelium prompting a strong pro-inflammatory response (Boulangé et al. 2016). Circulatory LPS has been linked to insulin resistance and obesity (Cani et al. 2007). Metabolites of the endocannabinoid system (ECS) were also negatively associated with the HMI. The ECS has been associated with metabolic syndrome (Bowles et al. 2015), IBD (Di Marzo and Izzo 2006), and coeliac disease (D'Argenio et al. 2007). Stimulation of the ECS can increase epithelial permeability (Muccioli et al. 2010), and a recent study showed the ECS has immunoregulatory roles in the gut (Acharya et al. 2017).

Overall, these results suggest a more disorder-associated gut microbiome might be associated with an increased intestinal permeability and host inflammation. However, these observations are largely speculative. They are based on broad level KEGG pathway annotations and only considered a smaller subset of individuals with existing metagenomic data. This reduced power might explain why observations were not observed between the HMI and other microbial functions known to modulate the host immune system, such as short-chain fatty acid production (Corrêa-Oliveira et al. 2016). Further mechanistic and intervention studies will be required to determine how differences in the functionality of the gut microbiota could contribute to multiple varied diseases.

6.4.5 Conclusion

In this chapter, I have presented a practical approach to identify marker taxa and reduce the data dimensionality within 16S rRNA gene sequencing-based gut microbiota profiles. This enabled uniform gut microbiota association analyses across numerous health effects. I found that microbiome features can have consistent, as opposed to disease specific, associations with multiple disorders. This provides a novel definition of a health-associated gut microbiota that can be quantified in the HMI. This index provides a robust method to query the gut microbiota in relation to host health and addresses several limitations

typically met using 16S rRNA gene sequencing approaches. The methods described in this chapter should be used to further understand the features that define a health-associated gut microbiome and how they can influence a range of diverse health effects. This will be key for the development of clinical interventions and diagnostics targeting gut microbiota.

Chapter 7

Identifying stable communities in the gut microbiota with consistent associations to host phenotypes

In this final research chapter, I present a data-driven method to identify robust communities of co-occurring gut microbiota within 16S rRNA gene sequencing datasets. I cluster reads from the combined data of three geographically diverse populations to create comparable OTUs. Applying the community detection approach to these data, I demonstrate the reproducibility of the method and show, for the first time, that gut microbiota form similar communities with similar host associations in different populations. These community level summaries provide a novel analytical unit that may be more representative of microbiota functionality than OTUs and also serve to reduce the dimensionality of 16S rRNA gene sequencing data.

Collaborator Attributions

Pre-processing of data from the Dutch Lifelines-DEEP cohort was carried out by collaborators at the University of Groningen, with replication of associations in these data carried out by Marc Jan Bonder.

7.1 Introduction

As discussed in Section 1.4.1, interactions between gut microbiota can influence the wider-scale composition and function of the gut microbiome. However, assessing interactions on a microbiota-wide scale is challenging and has only more recently been addressed by the development of methods to quantify co-occurrence between taxa (Section 1.4.2). As such, there has been little exploration of approaches to identify communities of interacting taxa in the gut microbiota.

Within a microbiome, individual taxa will not interact with all others. Rather they will form preferential interactions with subsets of taxa, due to factors such as cross-species metabolism and geospatial environmental variation (Faust and Raes 2012; Levy and Borenstein 2014). If such distinct communities of OTUs can be identified, they could represent more functionally relevant units for analysis than collapsed taxonomies. Furthermore, if such communities can be identified that are stable across populations and interact with host health, they could represent novel targets for gut microbiome-based medicines.

Interactions between OTUs can be inferred from the co-occurrence between them across multiple samples (Faust et al. 2012). However as OTU counts are relative to the sequencing depth of a sample, specialist approaches must be used that account for this when estimating correlations in microbiota data. As discussed in Section 1.4.2, a selection of such approaches were recently considered in a systematic comparison by Weiss *et al.* (Weiss et al. 2016). Comparing a range of approaches using several different quality measures, they found that using an ensemble of metrics can improve the precision of co-occurrence detection (Figure 1.10). Following the results of this study, an adaptation of the ensemble approach is used generate the interaction networks used in this chapter.

Previous studies have identified communities within microbial co-occurrence networks (Lozupone et al. 2012a; Tong et al. 2013; Duran-Pinedo et al. 2011), often using the WGCNA method as demonstrated in Chapter 3. However, whilst they may be more biologically accurate, such approaches allow OTUs to have weighted contributions to multiple communities. These overlapping definitions complicate comparative analyses and mapping of communities between networks. This can be overcome by using community detection algorithms that assign OTUs to single communities as discussed in Section 1.4.3. This was demonstrated in a study using Louvain modulation maximisation to compare community structures between irritable bowel syndrome patients and healthy controls (Baldassano and Bassett 2016). In this chapter I similarly use modularity maximisation to detect microbiota communities within OTU interaction networks.

In both generating co-occurrence networks and community detection, there are several stages at which thresholds or parameters must be selected. For instance, in relation to the present chapter, I must select a p-value threshold at which to consider correlation between two OTUs as an edge in the network and choose a value for a parameter (γ) in the

community detection algorithm. In this chapter I aim to avoid using arbitrary thresholds to define these, instead using data driven properties to identify optimal parameters.

To demonstrate the robustness of the community detection method used in this chapter I compare its output across three geographically diverse populations. To this end, I first generate comparable OTUs by combining the TwinsUK data with 16S rRNA gene sequencing-based gut microbiota profiles from two other cohorts from Israel and the Netherlands (Zeevi et al. 2015; Fu et al. 2015). I then apply a novel data-driven approach to generate co-occurrence networks and detect the communities within each dataset. This produces stable community definitions that can be compared and mapped between the datasets. I show that OTUs can form similar community structures with similar associations to host phenotypes across all three populations.

7.2 Methods

7.2.1 Data aggregation

Given the objective to compare network community structures across data sets (chapter overview in Figure 7.1), I required OTUs that would be comparable between them. To achieve this, I carried out de novo OTU clustering across combined sequencing data from multiple sources. To maximise sequencing similarity, I chose to use two data sets whose experimental approach best matched that used to generate the 16S rRNA gene sequencing profiles in TwinsUK. These were gut microbiota profiles from individuals in the Dutch Lifelines-DEEP (LLDEEP) cohort and from a personalised nutrition study within an Israeli population by Zeevi *et al.* (Israeli-PN) (Fu et al. 2015; Zeevi et al. 2015). All three data sets carried out gut microbiome profiling by amplifying the V4 region of the 16S rRNA gene using the same PCR primers, and used paired-end sequencing with read lengths sufficient to capture the whole V4 region. Notable differences between the studies include faecal sampling and DNA extraction techniques. Both TwinsUK and LLDEEP utilised aliquots from faecal samples stored at -80°C, the Israeli-PN study utilised a mixture of faecal swabs stored at -80°C and OMNIgene-GUT stool collection kits stored at -20°C. All three studies used both chemical and mechanical lysis in DNA extraction but employed different protocols: TwinsUK utilised the MoBio PowerSoil HTP extraction kit, the LLDEEP cohort utilised the Qiagen AllPrep kit, and the Israeli-PN DNA was extracted using the MoBio PowerMag Soil DNA extraction kit.

One sample per twin was selected from Batch 3 of the TwinsUK gut microbiota data (Table 2.1) as in Chapter 6, although the set was slightly larger as it was not restricted to those with accompanying disease phenotypes. The demultiplexed and merged paired-end data was provided by collaborators at Cornell University. Data from the LLDEEP cohort was provided in a similar format by collaborators at the University of Groningen, who used

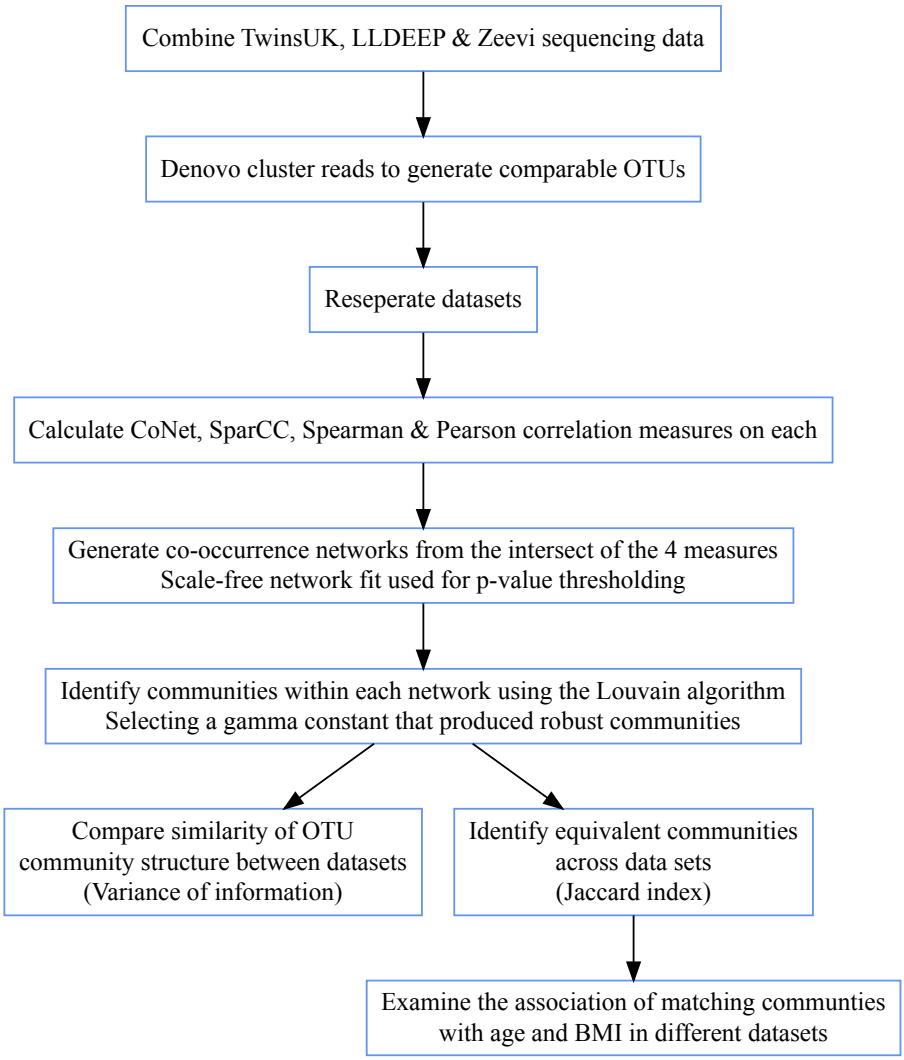


Figure 7.1: An outline of the study design used in this chapter.

custom scripts to merge the paired end data to full length reads covering the V4 region and split the data by individual (Gevers et al. 2014; Fu et al. 2015). Raw data accompanying the Israeli-PN publication was downloaded from the European Nucleotide Archive (ENA) (Accession:PRJEB11532) (Zeevi et al. 2015). This was processed similarly to the TwinsUK data and demultiplexed using published barcode mappings. Samples not listed in the accompanying metadata or with ambiguous read identifiers were removed. The cleaned, joined and demultiplexed data were then concatenated to produce one sequencing file covering all three studies. In total 4426 samples (2764 TwinsUK, 1023 LLDEEP, 639 Israeli-PN) were included in the analysis.

7.2.2 OTU clustering, table filtering, and diversity analyses

Generation of OTUs was carried out using the VSEARCH package, as this produced similar results to SUMACLUST in Chapter 5 but could better handle the large amount of data in this study (Rognes et al. 2016). The complete sequencing from all three sets contained 381,767,528 reads. These were dereplicated, removing reads occurring only once, resulting in 5,728,288 unique reads. De novo OTU clustering was carried out on the unique reads using cluster_fast with a 97% similarity threshold. The resultant 94,070 representative sequences were filtered to remove chimeric reads using the uchime_denovo command (within VSEARCH), producing a final set of 17,123 representative sequences. These were used to generate OTU counts across all samples using the VSEARCH usearch_global command. Post-processing, average read counts were 82,695±745 for TwinsUK, 49,962±964 for LLDEEP, and 130,378±5,534 for Israeli-PN (mean±SEM).

A phylogenetic tree was generated from representative sequences using the default parameters of the make_phylogeny command in QIIME. Taxonomy of OTUs was assigned by matching representative sequences against the Greengenes v13_8 database using the default parameters of the assign_taxonomy command in QIIME. OTUs occurring in only one sample, and samples with less than 10,000 OTU counts were removed. Weighted and unweighted UniFrac beta diversity measures and subsequent principle co-ordinates analysis of them was carried out using the beta_diversity_through_plots script in QIIME. The combined OTU table was then split by data set. For the purposes of alpha diversity calculations, the raw counts tables were rarefied to a depth of 10,000 reads. For each sample, Shannon and inverse Simpson diversity indices were calculated as the mean across ten rarefactions. Significant differences in alpha diversity between datasets were assessed using Mann-Whitney U-tests.

7.2.3 An ensemble approach to co-occurrence networks

Co-occurrence calculations were carried out on each dataset independently. Sub-tables were generated from raw (unrarefied) OTU tables that only contained OTUs found in at least 25% of the samples. All co-occurrence measures were calculated within these subsets as they are less sparse and hence more amenable to correlation measures (Weiss et al. 2016). The mean diversity was assessed on rarefied versions of the subset tables using the inverse Simpson index as for the full tables. OTU sparsity was assessed from the unrarefied sub-tables, also used for the co-occurrence calculations, using the biom summarize_table command. Following the recommendations of Weiss *et al.* (Figure 1.10), I used estimates of the mean inverse Simpson index (TwinsUK=20.2, LLDEEP=29.0, and Israeli-PN=13.1) and OTU table sparsity (0.49 in all three) to select an ensemble approach to co-occurrence detection that combines four different correlation measures: CoNet, SparCC, Pearson's, and Spearman's.

CoNet CoNet is itself an ensemble approach. The package allows a range of co-occurrence measures to be combined with several options for how to combine the weighting of edges and their p-values. CoNet addresses compositionality within the data using the ReBoot procedure, which involves permutation followed by re-normalisation of data. Calculating co-occurrence within these renormalized data allows assessment of the levels of correlation expected simply as a result of the compositionality within the data (Faust et al. 2012). For this study I used four measures of co-occurrence within CoNet: Spearman's and Pearson's correlations and Kullback-Leibler and Bray-Curtis distance measures (Faust et al. 2012). Initial correlation thresholds were selected for each of these measures that produced 2000 positive and 2000 negative edges concordant across the four metrics (Weiss et al. 2016); 1000 permutations were then used for re-normalisation to account for compositionality, and bootstrapping to identify edge p-values for each metric. The Simes method was chosen to merge p-values across edges by keeping the minimum. Final p-values were adjusted for multiple testing using the Benjamini-Hochberg FDR approach.

SparCC SparCC was developed to calculate correlations between OTU abundances in microbiome data whilst accounting for their inherent sparsity and compositionality (Friedman and Alm 2012). It uses the centred log-ratio transformation to address data compositionality (Figure 1.9). SparCC was used with default parameters to calculate correlations from the raw count OTU tables. The MakeBoostraps command was used to generate 100 bootstrapped tables, which were in turn used to calculate SparCC correlations. The bootstrapped correlations were then used with the PseudoPvals command to generate two-tailed p-values for the SparCC correlations from the true table.

Pearson's and Spearman's Correlations Pearson's and Spearman's correlation metrics do not take data compositionality into account, but were included to follow the approach outlined by Weiss *et al.* (Weiss et al. 2016). Both measures were calculated using relative abundance tables. The rcorr function from the Hmisc R package was used to calculate correlations and generate two-tailed p-values pairwise between all OTUs for both Pearson's and Spearman's measures (Harrell Jr and Dupont 2008). P-values were adjusted for multiple testing using the Bonferroni method in R.

For each of the four co-occurrence approaches, the output was converted into two simplified unweighted edge tables one detailing the direction of association (1 or -1) between OTUs and another detailing the bootstrapped/adjusted p-values.

7.2.4 Data-driven determination of a p-value threshold

Intersected networks were generated by combining the edge tables from the CoNet, SparCC, Pearson's, and Spearman's methods. This was done independently for each data set. Edges were retained in the combined network if the direction of co-occurrence matched and the edge p-values were below a given threshold in all four methods. This was carried out at multiple different p-value thresholds (0.05, then ranging in powers of ten from 0.01 to 10-8), to generate multiple intersect networks for each dataset with a gradient of stringency for edge inclusion. The resultant networks were then assessed to determine how well their structure fit that of a scale-free network to identify the optimum p-value threshold to use.

A scale-free network is a network whose node degree distribution follows a power law, i.e. there are few highly connected nodes and many more less connected nodes, and is the expected distribution of several biological networks (Jeong et al. 2000; Albert 2005). Fit to a scale free network was calculated by first extracting the degree distribution of a network using igraph (Csardi and Nepusz 2006). The fit of this distribution was then assessed using the scaleFreeFitIndex function from the WGCNA package in R (Langfelder and Horvath 2008). This provides the R2 of the fit to a scale-free model, which the creators of WGCNA suggest should be >0.8, and the slope of the fit, which they suggest should be close to -2 to indicate a good fit. The optimum p-value threshold was selected based on these criteria across all three data sets and the resultant intersect networks at this threshold used in all further analyses. Visualisation and generation of descriptive statistics from the networks was carried out using Gephi (Bastian, Heymann, and Jacomy 2009).

7.2.5 Detecting communities within co-occurrence networks

Between OTU adjacency matrices were generated that represented the final intersect networks for each dataset. Negative correlations represented a considerably small proportion

of edges (<1%) in all data sets and were removed to generate unsigned networks. Community detection was carried out on each dataset's network independently using the genLouvain 2.0 package within MATLAB (Mucha et al. 2010), which implements the Louvain approach to modularity maximisation (Blondel et al. 2008). This defines community partitions by assigning nodes to unique communities, then iteratively combines neighbouring nodes into communities if it results in an increase in modularity across the whole network (see Section 1.4.3). Modularity is a measure of the numbers of edges within communities relative to between communities and a higher value represents better community definition (Newman and Girvan 2004) (Figure 1.11).

Data-driven determination of γ in community detection

The genLouvain algorithm includes a γ parameter, which controls the size and numbers of detected communities (Mucha et al. 2010). A smaller γ value promotes the detection of small numbers of larger communities, while larger γ values promote the detection of high numbers of smaller communities. To find an optimal value for the γ parameter, I carried out community detection using a range of γ values for each dataset (0.1-1, increments of 0.01), and selected the optimal γ value based on the stability and significance of the resultant communities.

Stability To assess the stability of community definitions at each γ, I carried out community detection 25 times on the real network. I then carried out pairwise comparisons of the similarity of community clustering between the 25 runs. To assess similarity between two community groupings I used the normalised variation of information (variation of information divided by ln(number of nodes in the network)) as a measure of similarity between assignments. This ranges from 0 to 1. A high value for variation of information means OTUs are sub set more differently in the two partitions compared, whereas a value of zero means the two partitions are identical. In figures and text, 1-normalised variation of information is reported, so a higher value represents more similar community structure.

Significance To assess the significance of the community definitions, I also carried out community detection at each γ on 100 randomised networks whose nodes followed the same degree distribution as the real network (generated using the randomGraphFromDegreeSequence command in the Octave networks toolbox in MATLAB). I then compared the mean modularity of the 25 runs on the real network to the 100 randomised networks.

From observations of both the variation of information (stability) and modularity (significance) results I selected a suitable value for γ that produced both stable and significant community groupings. Once a suitable value for γ was identified, community detection was

repeated 100 times at this value, and the community definitions from the run producing the highest modularity were retained as the final community definitions.

7.2.6 Community properties and host associations

The relative abundance of each community in a sample was found by summing the read counts of its constituent OTUs and dividing by the total number of reads observed in the sample. Association between the mean abundance of an OTU in a dataset and the number of OTUs in its parent community was assessed using Spearman's correlation. Taxonomy within communities was investigated by counting the number of OTUs assigned to different taxa at each taxonomic level.

The heritability of TwinsUK communities was estimated using the mets package in R as used in Chapter 6. Log transformed relative abundances were used for each community (log10 with the addition of 10^{-6} to account for zeros). For each community ACE, CE, E, and AE models were fit using data from complete twin pairs. Co-variates included age, BMI, sample collection method, sex, and sequencing depth. For each community, the best fitting model was determined as that with the lowest Akaike information criterion value.

Association analyses between community abundances and BMI and age were carried out using log transformed relative abundances of the communities. These phenotypes were selected as they were also available in both the LLDEEP and Israeli-PN data sets. Linear models were fitted for each community using the lm function in R. Community abundance was the dependent variable with BMI, age, gender, and sequencing depth as independent variables, as this was the maximal covariate set available across all three datasets. The coefficient and significance of associations were extracted for BMI and age. P-values were adjusted for multiple testing using the FDR method in R.

7.2.7 Comparison of community structure between datasets

Comparing overall community structure between datasets

I carried out pairwise comparisons between datasets to assess the similarity of overall community structures between the networks. In each comparison, the two networks were subset to just the OTUs shared by both and the normalised variation of information between their communities calculated. To find the variation that would be expected by chance, I generated randomised community sets for each network by shuffling OTU labels. Each randomised comparison therefore shared the same number OTUs between the two real networks with the same community sizes, but without the biological basis for the linkage of OTUs. I then carried out pairwise comparisons of the variation of information between the

randomised communities. Shuffling and comparisons were repeated 1,000 times. The highest score observed (1- normalised variation of information) in any pairwise comparison, in any permutation was 0.41 (mean=0.34, SD=0.014).

Comparing individual communities between datasets

To map communities between networks I carried out pairwise comparisons between all the communities in all three networks. The Jaccard index was used to quantify the number of OTUs shared between the two communities relative to the number not shared in each comparison. Matches were considered positive with Jaccard scores >0.25 (range 0-1, no shared OTUs - complete overlap). This was chosen as above this threshold there were no instances of multiple mapping of communities between datasets.

Community-types were defined where matches could be found linking the communities in all three data sets. These were labelled with colour names using the standardColors function from the WGCNA package in R. The log transformed abundances for the 14 community-types that could be mapped across all three data sets were generated for the LLDEEP and Israeli-PN data sets as for TwinsUK. These were analysed for associations with age and BMI using linear regressions as for TwinsUK.

7.3 Results

7.3.1 Overlap of the TwinsUK, LLDEEP, and Israeli-PN data

Summaries of the TwinsUK, Dutch LLDEEP, and Israeli-PN studies whose data were combined to derive the OTUs in this analysis can be found in Table 7.1. Principle coordinates analysis of beta diversity between samples found that whilst there was some grouping by dataset, there was also significant overlap between them (Figure 7.2A). Comparison of alpha diversities between the populations found significant differences between all three (Figure 7.2B).

Table 7.1: Participant summary statistics for the datasets considered. Values are Mean±SD

Dataset	BMI	Age	Sex (M/F/Unknown)	No. Samples
TwinsUK	26.1±4.8	59.5±12.3	308 / 2456 / 0	2764
LLDEEP	25.3±4.2	45.3±13.7	445 / 578 / 0	1023
Israeli-PN	26.4±5	43.5±12.9	376 / 251 / 12	639

Across the 15361 OTUs detected in the complete data 48% were shared across all three data sets, with 79% being found in at least two datasets (Figure 7.3A). There was also considerable overlap between TwinsUK and LLDEEP of OTUs not found in the Israeli-PN

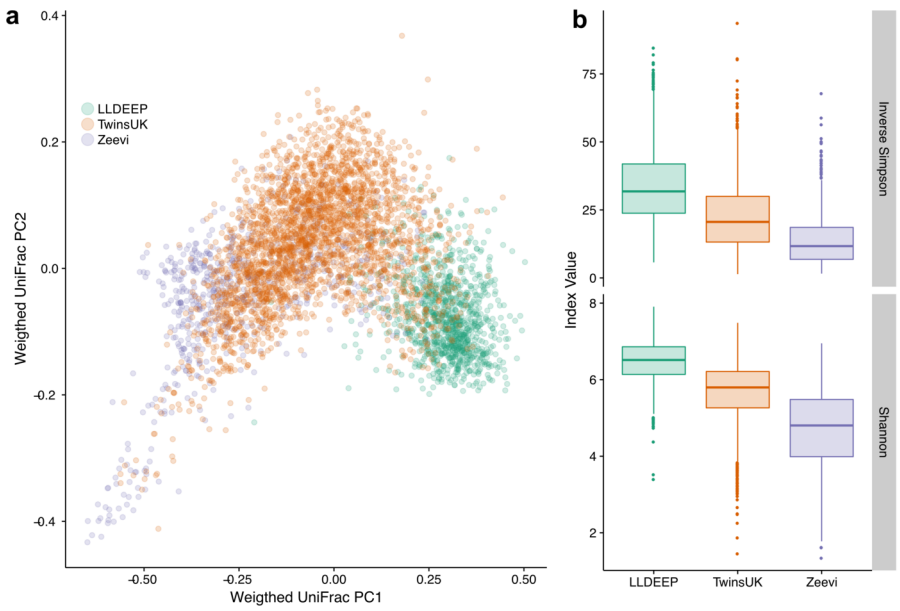


Figure 7.2: Diveristy comparisons between datasets. A) Plot of the first two components of PCoA from weighted UniFrac distance measures between samples in the study, coloured by dataset, shows there is some separation but significant overlap in microbiome compositions by study. B) Comparison of alpha diversity measures in all three data sets. There was a significant difference in all pair-wise comparisons for both measures (Mann-Whitney U $p<0.0001$).

data. Similar patterns were observed when considering more abundant OTUs (found in >25% of the respective populations). Although LLDEEP and TwinsUK shared more OTUs, examining the mean taxonomic distributions at the phyla level the TwinsUK and Israeli-PN cohorts were most alike (Figure 7.3B). The LLDEEP cohort contained relatively lower levels of the phylum Bacteroidetes, and a higher abundance of Firmicutes bacteria. Further differences were observed at the genus level, where the Israeli-PN study had a higher average abundance of Prevotella.

7.3.2 Scale-free thresholding of edges in ensemble networks

The datasets were split prior to generation of their co-occurrence networks using an ensemble approach - taking the intersect of four different correlation measures (TwinsUK example shown in Figure 7.4). Ensemble networks created using a range of p-value cut-offs were used to identify a final p-value threshold for edge inclusion, based on the networks' fit to a scale-free degree distribution (Figure 7.5).

There was no consistent trend across the p-values. This might be expected given the differences in the number of OTUs and samples in the datasets, and varying levels of p-value precision between the co-occurrence measures used. However, including all edges with a p-value <0.01 produced intersect networks with a good fit to the scale-free model in all three data sets. As such, this threshold was selected to generate the networks used for community detection.

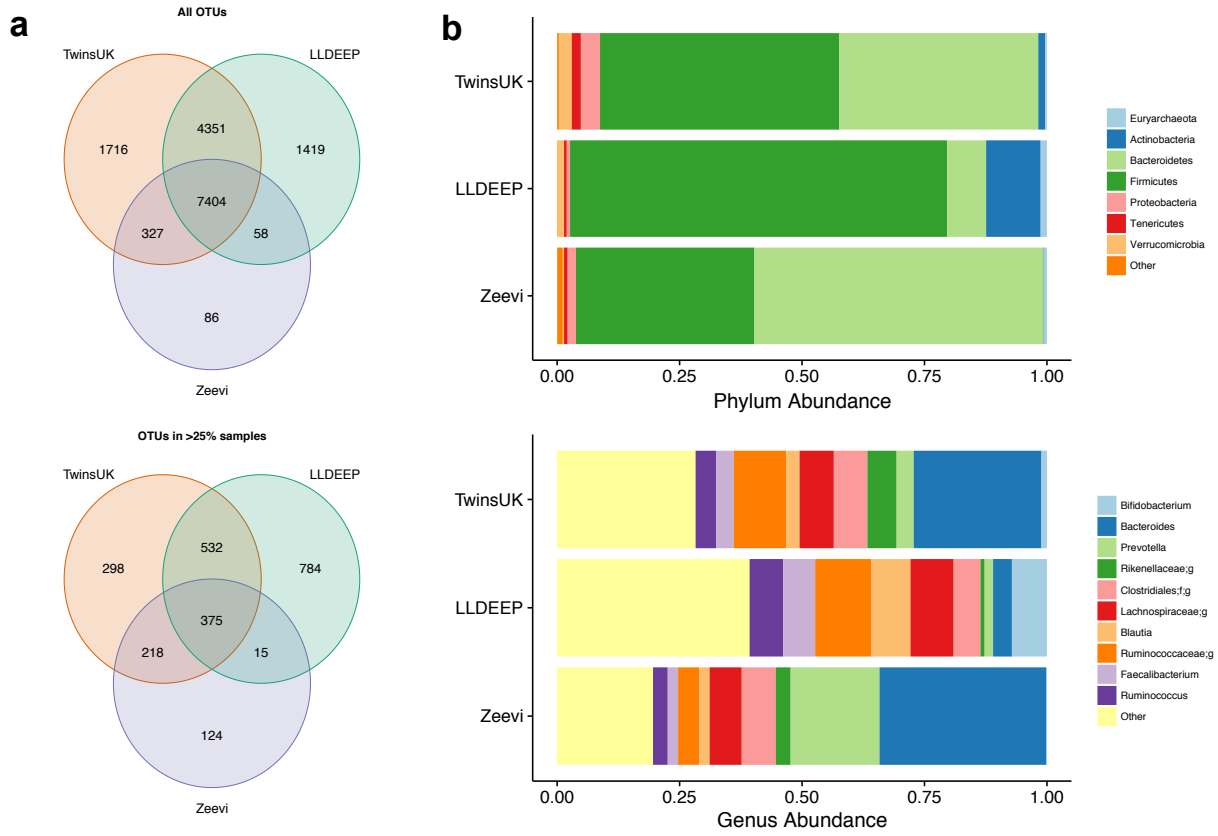


Figure 7.3: Taxonomic similarity of the datasets considered. A) Venn diagrams showing the number of OTUs shared across datasets. The top includes all OTUs, the lower diagram only those found in at least 25% of samples in each dataset. B) Comparison of the mean relative abundances of taxonomies at the phylum and genus level across the complete table for each data set. Phyla at less than 1% abundance and genera at less than 5% abundance in all sets are collapsed as Other. In genera f; and g; represent unassigned family and genus names in the Greengenes reference.

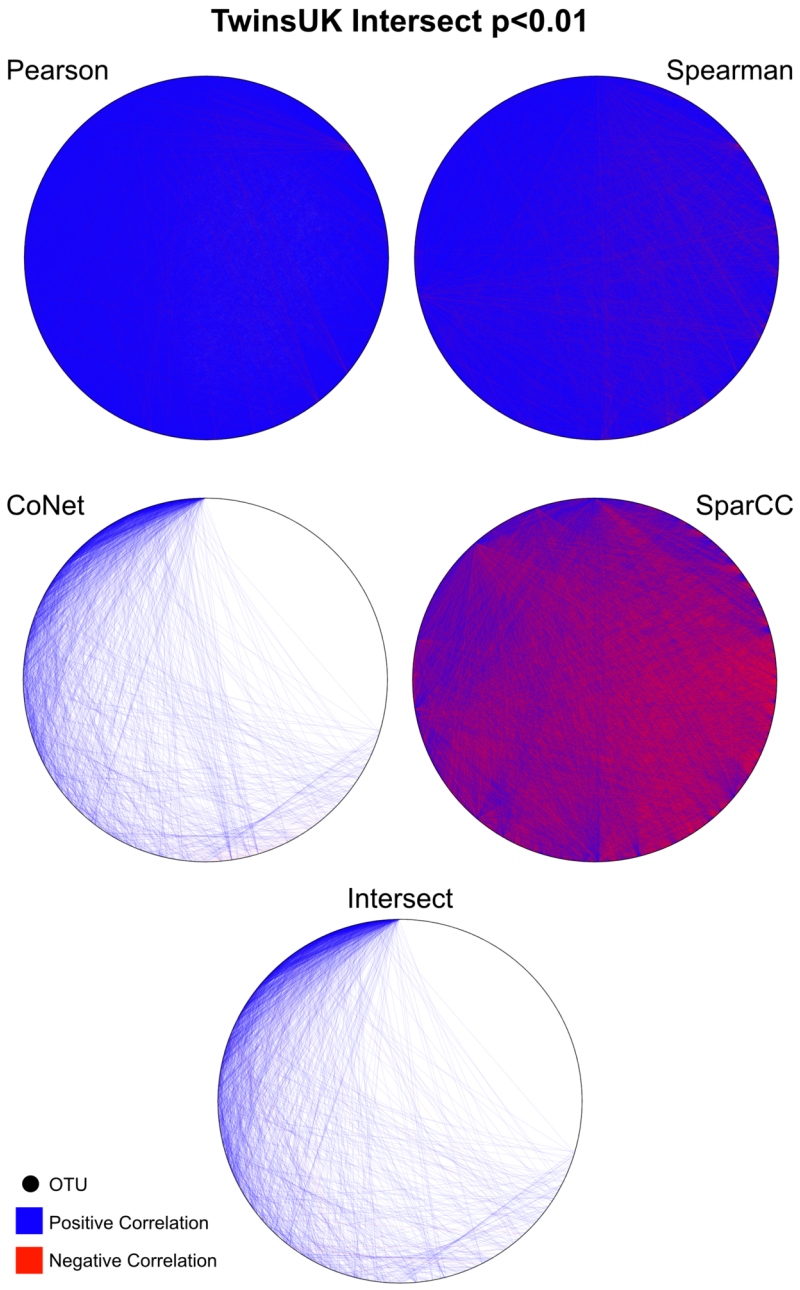


Figure 7.4: An example visualisation of the ensemble process at the p<0.01 threshold for TwinsUK. the top four networks combine to make the intersect network below.

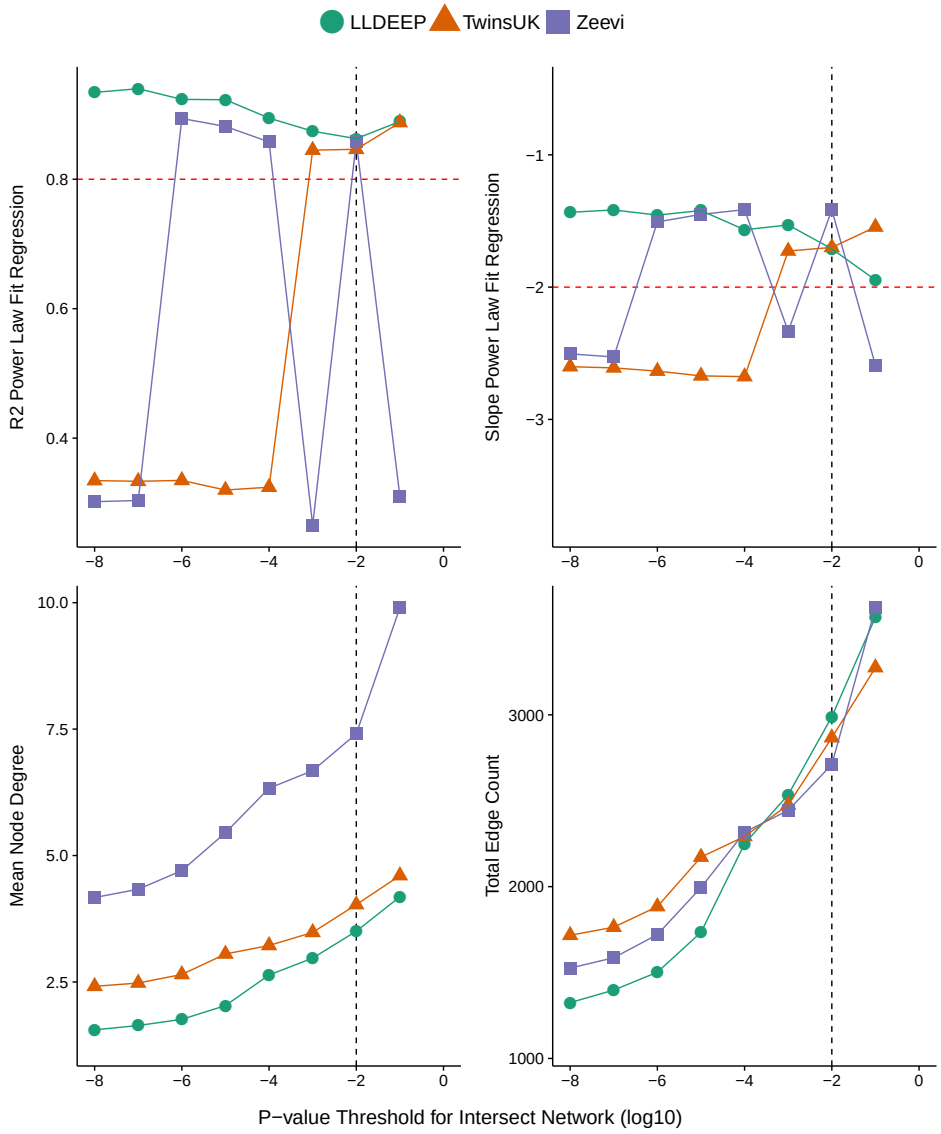


Figure 7.5: Selection of final p-value threshold for generating ensemble networks. Networks for each data set were made by intersecting four correlation measures with different p-value thresholds. Top left, the R2 of the resultant intersect networks fit to a scale-free distribution as assessed using WGCNA. Highlighted is 0.8 which the developers suggest as the minimum value to consider a good fit. Top right, the slope of the regression fit to a scale-free network returned by WGCNA, again highlighted is the developers recommended value. Bottom, summary measures of the resultant networks at each value showing the mean node degree of OTUs in the networks and the total number of edges in each.

From the 6761 unique edges observed across in the three final networks, 166 were observed in all three, 882 in the TwinsUK and LLDEEP, 677 in TwinsUK and Israeli-PN, and 229 in the LLDEEP and Israeli-PN datasets. Summary statistics for the resultant networks are shown in Table 7.2. The TwinsUK and LLDEEP network structures were most alike. The Israeli-PN data, which contained fewer nodes (OTUs), produced a smaller and more connected network.

Table 7.2: Summary statistics for the final intersect co-occurrence networks ($p<0.01$) for each dataset. Graph density is the percentage of all possible edges represented, mean path length is pairwise between all nodes, mean clustering coefficient is across all nodes in the network and provides a comparative indicator of overall clustering in the networks.

Dataset	Nodes	Edges	Mean Node Degree	Graph Density	Mean Path Length	Mean Clustering Coefficient
TwinsUK	844	2843	3.3	0.008	5.9	0.29
LLDEEP	922	2967	3.2	0.007	5.9	0.28
Israeli-PN	406	2573	6.7	0.033	3.7	0.31

7.3.3 Detection of communities in co-occurrence networks

After establishing co-occurrence networks I used the Louvain modularity maximisation algorithm to detect communities within them (Mucha et al. 2010; Blondel et al. 2008; Newman and Girvan 2004). The version used includes the constant γ that can be used to manipulate the number and size of resultant communities. To determine an appropriate value for γ, I carried out repeated community detection on the three co-occurrence networks at various γ values and assessed the stability (as variation of information between runs) and significance (mean modularity of real compared to randomised networks) of the community definitions at each γ (Figure 7.6).

Stability of community definitions

Community definitions were very stable (low variation of information between runs) at γ values approaching zero. This would be expected as in this instance most OTUs fall into few large communities. The mean variation of information fluctuated across the γ range. However, it should be noted that the lowest estimates (1-variation of information) were above 0.9 (range 0-1, least to most similarity in partitioning between runs), showing that even the least stable γ produced very similar communities between runs.

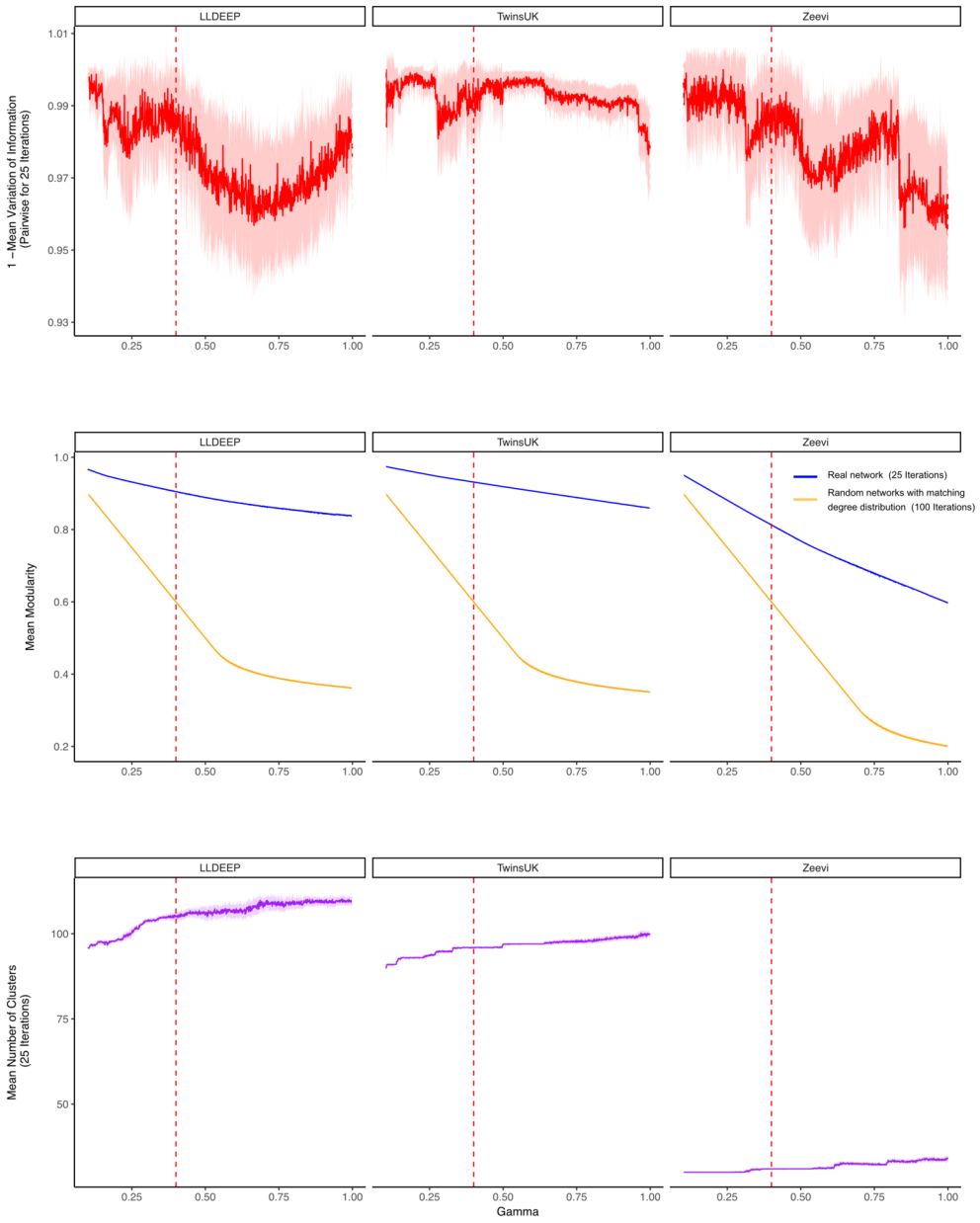


Figure 7.6: Selecting a γ parameter for Louvain thresholding. Louvain community detection was carried out multiple times on each network at a range of γ values. The top row shows the mean variation of information between the 25 runs at each γ value, a measure of stability of communities. Plotted is 1-variation of information so a higher value means more stable community definitions between runs. The middle row show the mean modularity of the 25 iterations compared to the mean modularity of 100 randomised networks with the same degree distribution, as a measure of clustering significance. The lowest row shows the mean number of clusters generated at each value of γ. In all cases the mean line is surrounded by the standard deviation across the interactions. This is also true for the middle row, however modularity estimates had extremely low variance between runs, hence even a small difference between the random and true means is significant. The dashed red line indicates the final selected value for γ.

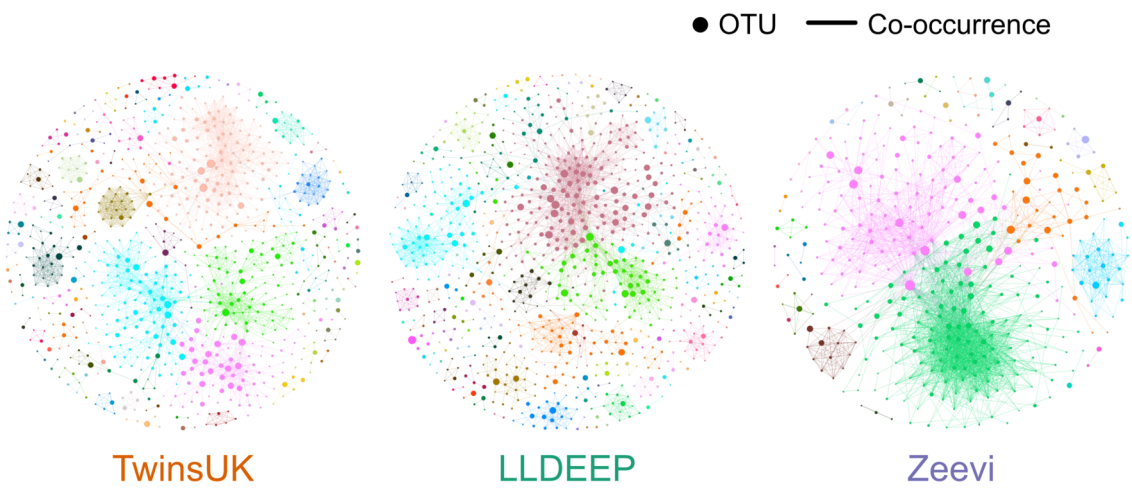


Figure 7.7: A visualisation of the communities detected by the Louvain algorithm. Communities are arbitrarily coloured. Node size represents an OTUs relative abundance within the dataset.

Significance of community definitions

At low γ values (with large communities) a networks edges are likely to be within communities, hence high modularity values were observed for both the true and randomised networks. As γ increased, modularity estimates for the random networks dropped significantly lower than those of the true networks (Figure 7.6). This shows that the real co-occurrence networks have more community structure than would be expected by chance, and provides confidence that the communities identified at these γ values reflect biological relationships between member OTUs.

Definition of final communities

Modularity and stability were highest in TwinsUK, followed by the LLDEEP, and then Israeli-PN data. This follows the sample sizes of the studies, and likely results from their comparative power to detect edges. From Figure 7.6, I selected a γ value of 0.4 for community detection. This γ provided high modularity estimates (that were significantly higher than the random networks) and good stability in all three data sets. The Louvain algorithm was then used with γ=0.4 to define the final OTU communities in the three co-occurrence networks (Figure 7.7). Communities in each network are hence referred to using arbitrary numbers ranging from 1 to the number of communities in the network.

The number of communities detected within the networks (TwinsUK=96, LLDEEP=105, Israeli-PN=31) reflected the number of OTUs within them, but was more similar when only considering communities containing at least 5 OTUs (TwinsUK=35, LLDEEP=36, Israeli-PN=11). There was a positive correlation between the mean relative abundance of an OTU and the total number of OTUs in the community it was assigned to in the TwinsUK and LLDEEP networks (ρ=0.1, p=0.002 and ρ=0.27, p<0.0001 respectively); i.e. high abundance OTUs were likely to be assigned to communities containing more OTUs.

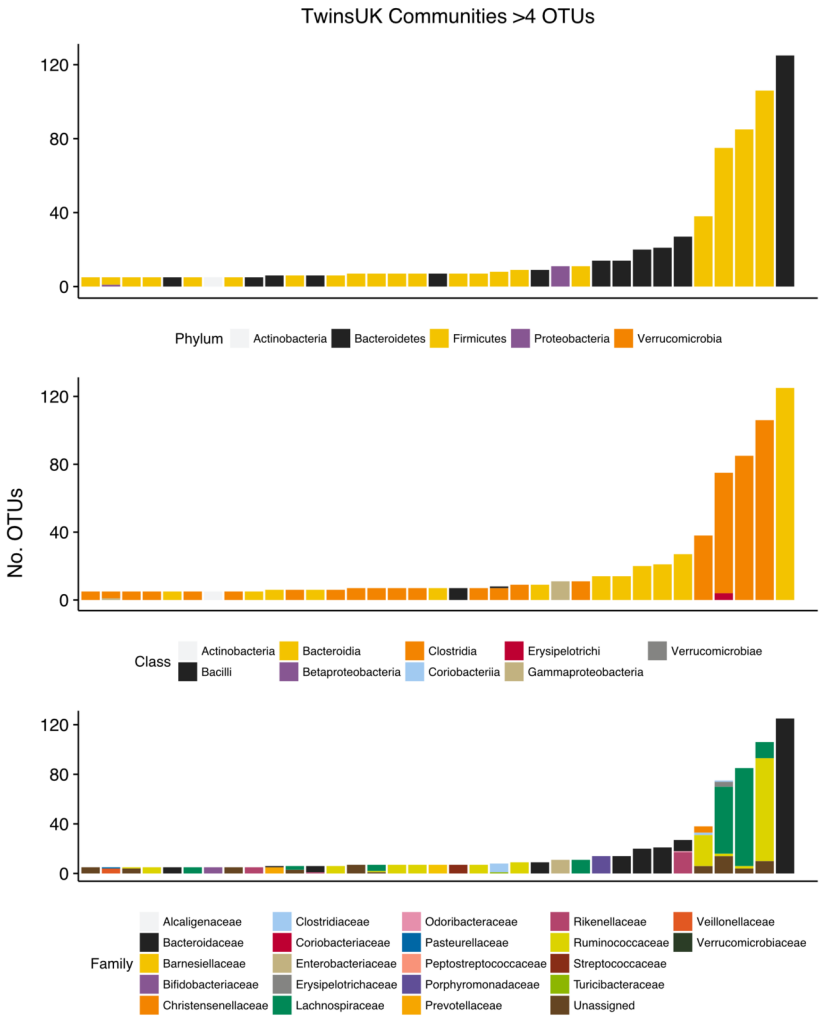


Figure 7.8: Summary of the taxonomic distributions within the TwinsUK communities. Each bar represents a community with communities ordered by size. Only communities with at least 5 OTUs are shown. Stacked bars represent the number of OTUs in the community assigned to each taxon.

However, all communities in all three networks contained a range of both high and low abundance OTUs (Figure E.1).

7.3.4 Taxonomy, heritability, and host associations of communities in TwinsUK

Investigating taxonomic assignments of OTUs within the TwinsUK communities, showed that at finer taxonomic levels (species and genus) several communities contained OTUs with a mixture of assignments (Table E.1). Whilst some communities retained a mixture of taxa at the family level, the majority were dominated by one taxon (Figure 7.8), and at higher levels nearly all contained a single taxon. A similar pattern was observed in the LLDEEP and Israeli-PN networks (Figure E.2).

ACE modelling was used to identify genetic influences on the microbial communities in TwinsUK within 654 MZ and 495 DZ pairs. From the 96 communities in the TwinsUK

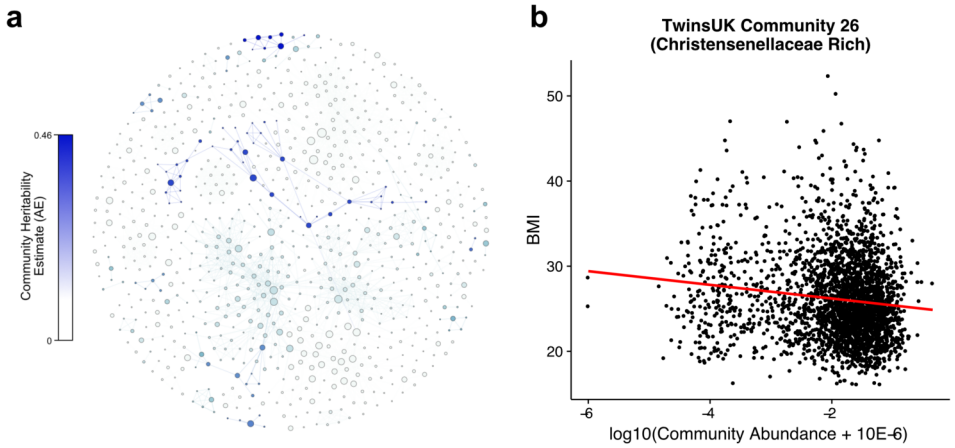


Figure 7.9: Host associations with TwinsUK communities. A) Community plot for TwinsUK as in Figure 7.7, but coloured by the heritability estimate of the community. B) Plot highlighting the significant negative association between BMI and the abundance of a Christensenellaceae rich community (linear regression with age, gender, and sequencing depth as covariates β=-0.13, FDR q=0.001).

network 52 had some variance attributed to additive genetic effects (Table E.1). Within these, the variance due to genetic effects ranged from 0.07-0.46 (Figure 7.9A). The estimates for the most heritable communities are of a similar magnitude to those of the most heritable taxa previously reported within TwinsUK (Goodrich et al. 2014b). Indeed, examination of taxonomies within these communities found that they contained heritable taxa. For instance, the most heritable community (Community 39) contained a Turicibacter OTU, and the second most (Community 26) contained numerous Christensenellaceae and Ruminococcaceae OTUs (Goodrich et al. 2014b).

Community associations with BMI

To further explore host effects, I carried out linear regression analyses between community abundances and age and BMI in TwinsUK (Table E.2). These phenotypes were chosen as they were available for all three datasets enabling replication. In TwinsUK, from the 96 communities 47 were significantly associated with BMI (FDR adjusted p <0.05). The highly heritable Community 26 that contained multiple Christensenellaceae OTUs had a significant negative correlation with BMI (FDR adjusted p<10-10, β=-0.13) (Figure 7.9B), reflecting previous observations in TwinsUK that Christensenellaceae and its correlated taxa are protective against weight gain in mice (Goodrich et al. 2014b). There were several communities containing exclusively OTUs belonging to the order Clostridiales that were negatively associated with BMI. The strongest positive association with BMI was with Community 5 (FDR adjusted p<10-7, β=0.11). This was a large community dominated by OTUs assigned to the Lachnospiraceae family.

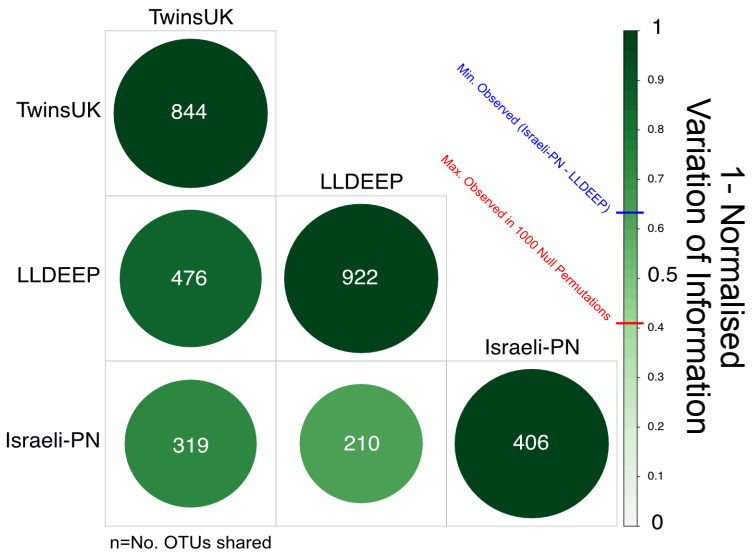


Figure 7.10: Pair-wise comparison of the variation of information between community definitions for OTUs shared between networks. 1-variation of information is shown where 1 represents identical segmentation of OTUs to communities and 0 indicates no similarity, with the highest score observed in randomised permutations being 0.41. Numbers represent the number of OTUs shared between the sets (OTUs with at least one edge).

Community associations with age

There were 48 communities significantly associated with age (FDR adjusted p <0.05). The three most significant associations were negative, two of these (Community 54 and 24), were dominated by Bifidobacterium OTUs. There was also a significant negative association with Community 27 (FDR adjusted p=0.007, β=-0.05), containing *F.prausnitzii* OTUs, replicating results observed in Chapter 3. The strongest positive association with age was observed with Community 17 (FDR adjusted p<10-7, β=0.11), which contained exclusively Enterobacteriaceae OTUs. There were also several small communities positively associated with age that contained exclusively OTUs assigned to the Ruminococcaceae family.

7.3.5 Community structure and associations across populations

Wide-scale overlap in community structure

Having identified communities within each dataset I aimed to determine how similar the segmentation of OTUs was between them. The normalised variation of information was again used to compare groupings. The similarity of community assignments was significantly higher in all three pair-wise comparisons than would be expected by chance (Figure 7.10). This shows that OTUs are forming similar communities in all three data sets. However, it should be noted that variation of information can only be used to compare matching sets, so this only shows that the OTUs shared between the populations have a similar community structure.

Mapping equivalent communities between datasets

I mapped the communities between the three networks using the Jaccard index to identify those which were equivalent to one another. Across the 96 TwinsUK, 105 LLDEEP, and 31 Israeli-PN communities, there were 14 instances where communities could be matched in all three data sets (Figure 7.11), although many more could be mapped across only two datasets. For distinction, I refer to these 14 matched groups as community-types and label each using arbitrary colour names (Figure 7.11 A & B, Table E.3). For example, the Green community-type was community 45 in the LLDEEP network, community 31 in the Israeli-PN network, and community 36 in the TwinsUK network.

Associations with age and BMI across populations

To determine if each of the 14 community-types had similar associations with age and BMI in their respective populations, linear regressions were carried out as for TwinsUK. Comparing the results (Figure 7.12), seven significant BMI associations within TwinsUK were replicated in the LLDEEP data, with two also significant in the Israeli-PN data. There were six community-types significantly associated with age in both the TwinsUK and LLDEEP data, four of which were also significant in the Israeli-PN data. BMI had more negative associations within the community-types, whereas age had more positive associations. For both BMI and age, the significance of community associations was most similar between TwinsUK and LLDEEP, this is may be due to their higher samples sizes relative to the Israeli-PN dataset.

The Greenyellow and Green community-types had significant negative associations with BMI in all three datasets. Greenyellow communities were dominated by Ruminococcus OTUs in all three, and similarly all the Green communities consisted solely of Ruminococcaceae OTUs. More widely, most community-types with at least one significant negative association with BMI consisted of Clostridiales OTUs. The Pink community-type for instance had a significant negative association with BMI in TwinsUK and LLDEEP, and contained multiple Lachnospiraceae OTUs assigned to the species *Coprococcus eutactus* in all three data sets.

The Salmon, Turquoise, Cyan, and Red community-types were significantly positively associated with age across all three data sets. All three contained exclusively Clostridiales OTUs. The strongest positive association with age in TwinsUK was with the Brown module which was not associated with age in the other sets, this contained Veillonella and *Haemophilus parainfluenzae* OTUs. Several associations for both age and BMI were significant in TwinsUK and LLDEEP but not in the smaller Israeli-PN dataset, but tended to share the same direction of effect. Overall, these results show that stable gut microbiota communities can be detected across populations and that they can maintain similar associations with host factors such as age and BMI.

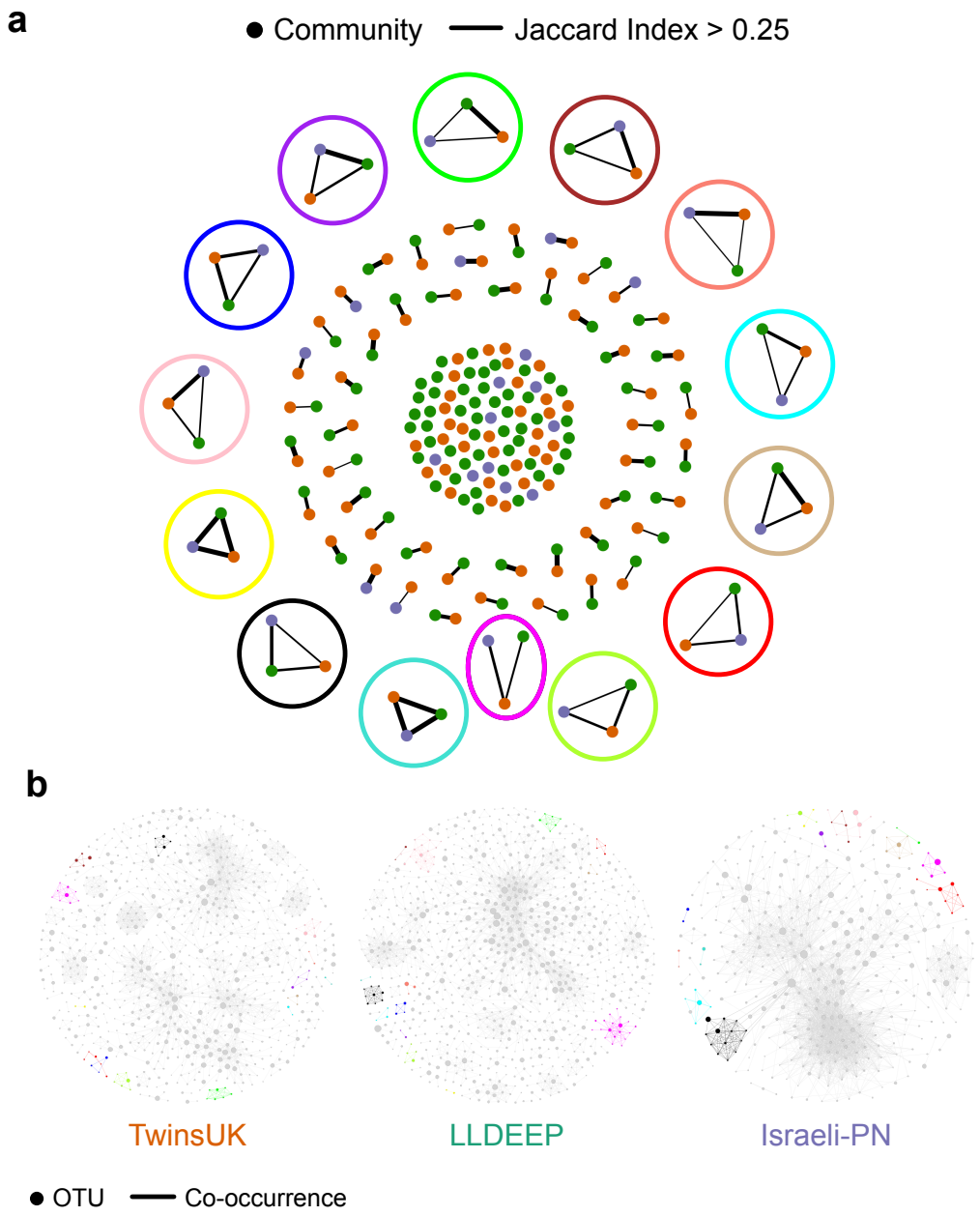


Figure 7.11: Mapping communities across datasets to identify stable community-types. A) A network showing the matching of communities based on Jaccard index. Edges are considered where the index >0.25 and are weighted by the index. Highlighted are the community-types selected for later analysis, using the colours matching their given names. B) Visualisation of networks as in Figure 7.7 but coloured based on community-types as in A. Communities not in these 14 types are coloured grey.

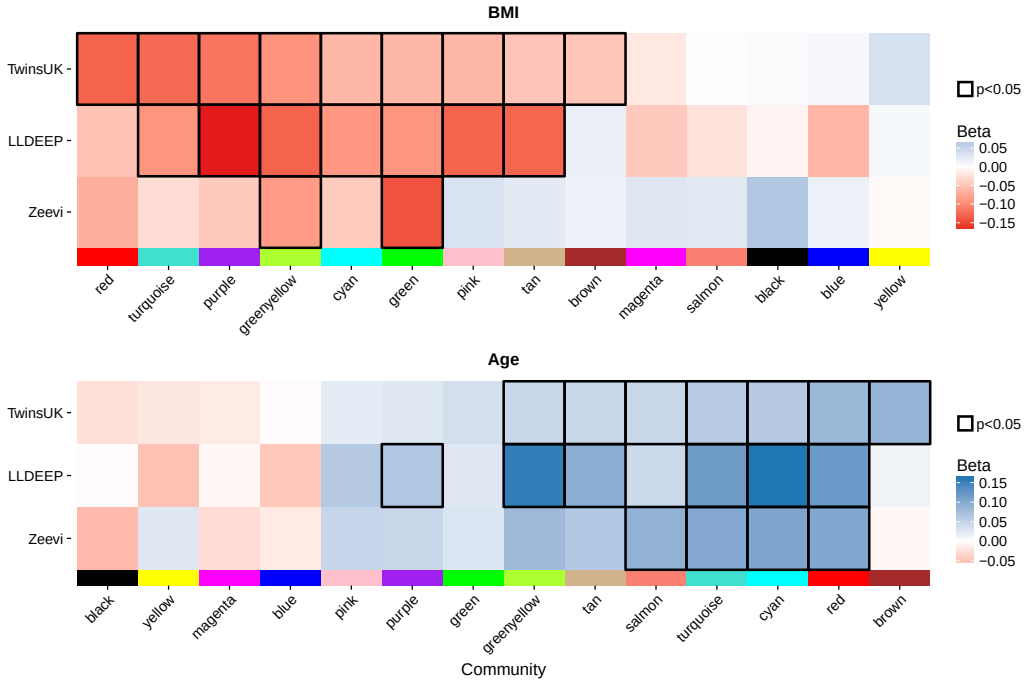


Figure 7.12: Replication of linear regression analysis for each of the 14 mapped community-types in the LLDEEP and Israeli-PN datasets. Top comparing the results for BMI and below for age. Squares are coloured by the Beta estimate for the association and nominally significant associations (p<0.05) are highlighted by a black border. Community-types are ordered by their associations within TwinsUK.

7.4 Discussion

In this chapter I have presented a rationalised and data-driven approach to generate comparable co-occurrence networks from 16S rRNA gene sequencing-based microbiota data and identify community structures within them. Applying this to gut microbiome profiles from three different populations showed that both co-occurrence networks and community structures are stable across all three. Furthermore, host factors such as genetics, age and BMI can influence the relative abundances of the identified communities, with age and BMI having similar associations with community-types mapped across the populations. The communities defined using this approach provide a further method to reduce the dimensionality of 16S rRNA gene sequencing profiles and could prove to be more functionally relevant analytical units than OTUs or collapsed taxonomies.

7.4.1 Differences in the gut microbiota between datasets

Significant differences were observed in the diversity and taxonomic profiles of the three datasets used in this chapter. These likely reflect a mixture of both differences in the study populations and the experimental approaches used. There will be differences between cohorts in factors known to influence the gut microbiota, such as genetics, diet and geography (Goodrich et al. 2014b; David et al. 2014; Yatsunenko et al. 2012). For instance, low diversity Prevotella dominant microbiomes have been previously linked to dietary intake (Wu et al.

2011). The larger proportion of Prevotella dominance, and associated reduced diversity, in the Israeli-PN gut microbiomes might therefore reflect dietary differences. However, technical disparities such as faecal sampling and extraction method, which were different between the studies, can also influence taxonomic composition and cannot be discounted (Walker et al. 2015). The increased overlap in the TwinsUK and LLDEEP data in comparison to the Israeli-PN dataset in both the OTU and network structures might also result from the higher sample sizes in these studies. However, it is notable that even with the intrinsic population and technical differences between the three datasets there were still consistent elements across them. Most OTUs were found in at least two studies and there was overlap of samples from all three in beta diversity PCoA plots. It is within this common ground that similarities in community structure were observed.

7.4.2 Generating interaction networks

To generate the co-occurrence networks in this study I used an ensemble approach as suggested by Weiss *et al.* (Weiss et al. 2016). However, CoNet was the main edge filter when intersecting networks and might alone be sufficient. This might be expected given that it is itself an ensemble approach. However, metrics such as SparCC are not implemented within CoNet, and different correlation methods can produce better results depending on the properties of the input dataset (Weiss et al. 2016). Therefore these observations may only apply to the current study and would require further testing before application to other datasets.

I extended the ensemble approach by selecting edges for inclusion in the interaction networks based on their fit to a scale-free topology. It has previously been shown that microbial interaction network structures can approximate a scale-free distribution (Chaffron et al. 2010). Scale-free networks are also well conserved in other aspects of biology across different domains of life (Albert 2005; Wolf, Karev, and Koonin 2002), and this approach is used by existing methods such as WGCNA (Langfelder and Horvath 2008). The observed stability of the resultant communities when using this approach indicates that using scale-free thresholding can provide analytical benefits by generating comparable network structures prior to community detection.

7.4.3 Robust community detection

Following network creation, I used modularity maximisation to assign OTUs to single communities. This may not represent the true nature of interactions, as it is likely that taxa contribute in different levels to different communities. Whilst it is possible to compare overlapping community definitions using more complex methods provided by packages such as WGCNA (Langfelder and Horvath 2008); using non-overlapping community def-

initions provides analytical benefits. It enables the use of simpler comparative measures such as the Jaccard index, which simplify analyses and the interpretation of results. There are also applications where unambiguous community definitions are desirable. For instance, in the design of synthetic therapeutic communities based on experimental observations, a method which could have clinical applications (Lindemann et al. 2016).

The community detection approach described here could be applied to investigate community differences between disease and control groups (Baldassano and Bassett 2016) or, as demonstrated, in replication of community associations across datasets. The importance of further development of reproducible methods to detect and analyse microbiota communities is highlighted by several studies that have shown associations between taxa in the gut microbiome and host health can depend on wider community context (Gevers et al. 2014; Goodrich et al. 2014b; Baldassano and Bassett 2016; Ridaura et al. 2013).

Beyond simply identifying communities, understanding the factors driving their formation would require extension of the described approaches to quantification of taxa using metagenomic sequencing. This, especially used in tandem with techniques such as metatranscriptomics and faecal metabolomics, might provide indications to the mechanisms of interaction. However, cross-sectional approaches are inherently limited to inference of interactions from co-occurrence across samples. Time-series data and in vitro studies will also be required to delineate directional effects and validate individual interactions (Faust et al. 2015). For instance, interactions validated in pair-wise species co-cultures can be used to predict outcomes in more complex multi-species cultures (Lindemann et al. 2016).

7.4.4 Stable community structure across populations

Although 16S rRNA gene sequencing cannot elucidate the mechanisms of the observed interactions, the comparisons within this chapter were sufficient to uncover novel biological phenomena. Most notably, that OTUs formed similar communities with similar host associations in the three different populations. This shows that the variation that exists between the populations (e.g. the OTUs unique to each dataset) does not significantly alter the interactions between the OTUs shared across all three.

Several community associations with age and BMI in TwinsUK replicated in the LLDEEP and Israeli-PN datasets. These associations, and the heritability results within TwinsUK, broadly and reassuringly reflected observations from studies investigating taxa in isolation. For instance, in TwinsUK, communities negatively associated with BMI were enriched with the butyrate producers Ruminococcaceae and *Coprococcus eutactus* (Louis and Flint 2009). Butyrate producing taxa in the gut microbiome have previously been associated with obesity, although at a higher abundance in obese mice (Turnbaugh et al. 2006). This may reflect an increased capacity for energy utilisation by the gut microbiota in leaner individuals, as butyrate has also been shown to have protective effects against metabolic

deficits associated with obesity (Qin et al. 2012; Gao et al. 2009). Two Bifidobacterium rich communities were negatively associated with age, also in agreement with previous studies (Yatsunenko et al. 2012; Odamaki et al. 2016), and *F.prausnitzii* and Enterobacteriaceae community associations with age reflected results from the frailty analyses in Chapter 3.

It would be expected that community level associations reflected those of taxa based studies as each community constituted largely of taxonomically similar OTUs. This is in agreement with a previous study that found co-occurrence to be higher between genetically similar taxa (Chaffron et al. 2010). This indicates either that interactions might evolve within closely related taxa, or that co-occurrence mainly detects genetically related taxa (more likely to have similar functionality) responding to environmental stimuli in the same manner. Either way, co-occurrence communities provide a useful method to reduce the dimensionality of 16S rRNA gene sequencing data to units reflecting biological phenomena. Further analyses using the described approaches should investigate the influence of host factors such as genetics, diet and health, on the formation of these communities and their reciprocal influences on the host.

7.4.5 Conclusion

In this chapter I have described a data-driven approach to identify comparable communities within microbiota co-occurrence networks. These communities provide a novel method to reduce the dimensionality of 16S rRNA gene sequencing data sets to units reflecting the biological properties of microbiota. This should facilitate future comparative, community-centric analyses of gut microbiome data. This utility was demonstrated in the dataset comparisons, which found that gut microbiota were able to form stable communities with similar associations to host phenotypes across populations.

Chapter 8

General Discussion

The aim of this thesis was to develop novel approaches to analyse gut microbiota profiles derived from 16S rRNA gene sequencing to improve their relevance to human health research. With a particular focus on addressing the high-dimensionality, limited reproducibility, and biological relevance of OTUs. This began with a discussion of the current state of the field in Chapter 1. Then in Chapters 3 and 4 I used simple approaches to investigate gut microbiota associations with frailty and PPI use. Chapter 5 described a comparative analysis to identify optimal OTU clustering parameters, whose results informed the methods used generate OTUs in latter analyses. Finally, in Chapters 6 and 7 I presented two new methods to summarise 16S rRNA gene sequencing data. The first in the health-associated microbiome definition and its subsequent quantification in the HMI; and the second in the detection and summation of robust interacting communities in the gut microbiota. In this concluding chapter, I provide a more detailed summary of these results, how they address the aims set out in Chapter 1, and their context in terms of one another and the wider field. I then go on to discuss some of the limitations of the presented work and how they might be addressed in the future.

8.1 Summary of results

8.1.1 Methodological Findings

Methods used in Chapters 3 and 4

Within the frailty study in Chapter 3, I used the Shannon index as a covariate to identify microbiota associations independent of wide-scale changes in microbiota diversity. To my knowledge this is a novel approach, and proved to be an effective way to identify the strongest associations. As such, it was similarly used in the gut microbiota association analyses with PPI use in Chapter 4. Wider adoption of this approach might be particularly

beneficial in gut microbiota studies aiming to identify taxa specific associations with health factors.

In Chapter 4, I presented an original approach to explore the mechanisms underlying the observed microbiota associations with PPI use. Using the public HMP data I inferred the body site from which associated taxa might have originated. In my analysis this was a largely speculative endeavour. However, a similar study investigating the effects of PPI use on the gut microbiota was published at the same time by researchers at the University of Groningen (Imhann et al. 2016). This study included oral microbiome sampling and confirmed the increased abundance of oral bacteria in the gut with PPI use. Ideally, if I were to repeat the PPI analyses, I would also collect oral microbiome data to confirm these effects. However, the broad agreement between the results in my analyses and those of the Groningen study lends support to the classifications created from the HMP data. Similar use of body-site classifications in other studies could enable inference of the mechanisms underlying associations where multi-site effects are involved.

Heritability as a quality measure

Chapter 5 described a comparison of different algorithms and parameters used to cluster 16S rRNA gene reads to OTUs. A key methodological finding in this chapter was the large influence that clustering approach can have on the relative alpha diversity estimates of samples. This has consequences for the interpretation and replication of results between studies using different clustering algorithms and diversity measures.

The most novel aspect of this chapter was the use of heritability as a quality metric. This was used to address the limited biological representation of OTUs: assuming that the clustering algorithm producing OTUs most representative of the units interacting with the host would produce the highest heritability estimates. This was largely motivated by the lack of a directly comparable gold-standard to enumerate the functional units in the gut microbiota. As such, heritability might similarly be applied as a quality measure in other fields lacking such standards.

Microbiota markers defining the HMI

The analyses in Chapter 6 built on the experience gained in the frailty and PPI analyses, extending similar approaches to multiple disorders and medications. There were two main methodological developments within this chapter. The use of marker traits and the development of the HMI.

The novel approach used to identify marker traits in the gut microbiome addressed the high-dimensionality of 16S rRNA gene sequencing datasets. A correlation-based approach to select markers was previously used in a large study investigating the relative influence

of host phenotypes on the gut microbiota within the Flemish Gut cohort (Falony et al. 2016). However, this was used to subset the phenotypes for consideration (rather than the microbiota traits) and did not use a network-based approach to select the optimal markers. Using a network allowed selection of a minimal set of taxa that represented the entirety of the wider set. This approach proved effective, identifying marker traits that eventually produced the HMI which accurately represented wider compositional patterns. Given the high levels of inter-correlation between different taxonomic summary levels in microbiome data, this approach could provide an effective way to select a minimal set of traits for consideration in microbiota from other body sites or even in environmental microbiome studies.

The definition of the health-associated microbiome and subsequently the HMI was made possible by analysing multiple diseases within the same cohort, overcoming the limited reproducibility typically observed between studies using 16S rRNA gene sequencing (Sze and Schloss 2016). The HMI reliably associated with microbiota composition and host health across different experimental populations and approaches. Quantifying microbiota composition in a single measure with directionality in relation to host health, addresses both the limitations of high-dimensionality and biological interpretation of gut microbiota profiles derived from 16S rRNA gene sequencing.

To date, microbiota composition has typically been quantified in alpha diversity metrics, under the assumption that a diverse microbiome is associated with health (Lozupone et al. 2012b); or using relative abundances of disease specific marker taxa, such as the MDI for IBD and the Bacteroides/Firmicutes ratio that was previously proposed to be associated with obesity (Gevers et al. 2014; Ley et al. 2006). The HMI is an extension and generalisation of the latter approaches. Previous efforts to describe a general healthy gut microbiome have relied on use of a screened healthy control population (Consortium 2012; Zhernakova et al. 2016; Falony et al. 2016). For example, a method was recently proposed in which healthy samples are used to define a 'healthy plane' within multidimensional space (Halfvarson et al. 2017). The dysbiosis of a disease patients microbiota can then be quantified by its distance from this plane. However, this requires consistent and comparable enumeration of microbiota between all samples and assumes that perfect screening of a health can be achieved.

The HMI is defined using the definition of health as a contrast to disease. As such, the health and disorder definitions include the average observations across a range of different health and disorder states. This, and the broad familial level of its definition, are responsible for its applicability and utility across platforms, addressing the limitations of reproducibility in 16S rRNA gene sequencing-based microbiota studies. The HMI is one of the key outputs of this thesis that should influence future gut microbiome research. It provides a simple and accessible measure of microbiota composition that can also be used to determine if microbiota differences might be beneficial or detrimental to the host. This could have a

range of applications as discussed previously in Chapter 6.

Defining communities in microbiota networks

The final method described in this thesis was the approach to identify communities of interacting microbiota in Chapter 7. This used an OTU clustering technique identified in Chapter 5 and built on co-occurrence community ideas first explored in the frailty analyses of Chapter 3. The method used the inherent structure in the data to inform the parameters used in network creation and community detection. This resulted in community definitions that were stable across different populations. As such, the community detection approach plays a role in addressing all of the specific aims of this thesis; by providing a reproducible method to reduce the dimensionality of 16S rRNA gene sequencing datasets to units defined on a biological basis. This method could have wide applicability in future analyses investigating community level effects in the gut microbiota, which has been the aim of several previous studies (Baldassano and Bassett 2016; Tong et al. 2013; Duran-Pinedo et al. 2011). This might also be a particularly useful tool in the design of beneficial bacterial communities for therapeutic purposes. For instance creating synthetic communities to confer benefits such as those observed with faecal microbiota transplants (Vos 2013; De Roy et al. 2014). The approach described in Chapter 7 could enable identification of the communities responsible for beneficial effects within whole microbiome samples, aiding rational design of synthetic alternatives to untargeted treatments.

8.1.2 Biological Findings

Whilst the principle concern of this thesis was methods development, several novel biological phenomena were also observed through the application of the described approaches.

Generality of microbiota disease associations

I consider the most important biological observation of this thesis, in terms of contribution to the field, to be the demonstration that microbiota associations can be consistent across disorders, rather than being solely disease specific. This is not an entirely novel concept, reviews have previously aimed to identify similar patterns by condensing results across studies (Lloyd-Price, Abu-Ali, and Huttenhower 2016), but it is to my knowledge the first uniform experimental demonstration of the fact. This observation is clinically important as it indicates that there may be a common basis for some of the observed health associations with the gut microbiome; which in turn means that advances in gut microbiota mediated treatment of one disease might be more widely applicable. Furthermore, identification of a common pattern of associations is the first step towards understanding their underlying mechanisms.

Immune mediation of gut microbiota influences on host health

The analysis in Chapter 3 was not the first study to explore gut microbiota associations with host frailty. Whilst it identified some novel associations, such as those with *E.lenta* and *E.dolichum*, the strongest associations replicated existing results. These included the reduced alpha diversity and abundance of *F.prausnitzii* with frailty, which replicated observations from other studies of frailty and extremes of ageing (Odamaki et al. 2016; Claesson et al. 2012; Tongeren et al. 2005; Biagi et al. 2010).

The reduced abundance of butyrate producing *F.prausnitzii* in the gut microbiota has been associated with multiple diseases (Sokol et al. 2008; Song et al. 2016; Qin et al. 2014). Amongst other effects, microbial butyrate production can maintain immune homoeostasis in the gut (Corrêa-Oliveira et al. 2016). As frailty is associated with chronic low level inflammation (Hubbard and Woodhouse 2010), its associations with gut microbiota may be mediated by the immune system. This was supported by the results of Chapter 6, where the metagenomic and metabolomic enrichment analyses identified several potential pro-inflammatory microbial functions/faecal metabolites that were enriched in samples from individuals with a low HMI (or a more disorder associated gut microbiome). Combined these results lend support to an increasing body of work showing that immune modulation could be a major mechanism by which gut microbiota influence wide-scale health (Round and Mazmanian 2009). This was demonstrated in a recent study, which showed that germ-free mice can out-live mice with conventional gut microbiota and display reduced inflammatory markers across their lifespan (Thevaranjan et al. 2017). This concept has been further refined in the 'common-ground' hypothesis.

The 'common-ground' hypothesis was proposed in a recent review of the effects of gut microbiota in human health (Lynch and Pedersen 2016). It states that pro-inflammatory actions of the gut microbiota could promote inherent polygenic susceptibility to various diseases in the host, a hypothesis which fits the observations of Chapter 6. Modulation of host inflammation may be one mechanism by which similar changes in the gut microbiota are able to lead to a diverse range of health deficits. However, the observational nature of the studies in this thesis mean it is not possible to determine the order of causality. Differences in gut microbiota could also simply be indicators responding to inflammation from another, non-microbial, source. In particular, it is possible that these effects could largely be driven by changes in host health/immunity that alter the gut environmental niche and hence the microbes that inhabit it. Further research is required to determine if/how similar changes in the gut microbiota are able to influence a diverse range of health effects. However, the potential malleability of the gut microbiome makes it an enticing therapeutic target to promote overall health.

The influence of medications on the gut microbiota

In both Chapters 4 and 6, I considered the influence of medication use on the gut microbiota. The effects of PPI usage have been investigated previously but in smaller scale studies (Freedberg et al. 2015; Seto et al. 2014). As mentioned in Section 8.1.1, a similar population based cohort study was published at the same time and found almost identical associations including an increased abundance of oral bacteria, in particular Streptococcaceae, with PPI use (Imhann et al. 2016). In Chapter 6, I described associations with a wide-range of other medications the majority of which had not been considered in relation to the gut microbiota.

Bar a few exceptions (such as the metformin study discussed in Section 1.2.4), most investigations of drug influences on the gut microbiome have focused on the effects of antibiotics (Jernberg et al. 2007; Jakobsson et al. 2010; Falony et al. 2016). The results of Chapters 4 and 6 show that other, similarly common, medications can also have significant associations with gut microbiota. This has two main implications: Firstly, more drugs should be considered in screening of participants and as covariates in gut microbiota studies (discussed further in Section 8.2.1). Secondly, these results warrant further research into the clinical consequences of these interactions. For example, the associations with PPI use have been proposed to mediate an increased susceptibility to *Clostridium difficile* infection (Seto et al. 2014), and it has been posited that microbiota alterations may, in part, be responsible for the beneficial effects of metformin treatment (Shin et al. 2013).

Stable gut microbiota community configurations across populations

The final novel biological observation, was the consistency of the gut microbiota communities and their host associations across different populations in Chapter 7. Evidence that community structure is conserved is promising. It demonstrates that communities associated with health effects in one population might be more widely applicable, motivating further research to this effect.

8.2 Limitations and Future Work

8.2.1 Use of observational cohort data

Data quality

The wide range of data available in TwinsUK enabled the comparative analysis in Chapter 6. It also facilitated the quantification of covariates known to influence the gut microbiome

such as diet, which are often overlooked in less comprehensive studies. However, the TwinsUK data have limitations. Variables have been collected at a range of different times, using different questions or techniques, and not always at the same time as the faecal samples. This adds variability to the phenotypes that, especially given the temporal variation of the microbiome, increases the noise in the data and reduces the power to detect associations. Future work to address this would require collection of all phenotypes at the same time as microbiota sampling, ensuring all relevant data is collected. This would be a large effort and require many twin visits at great expense to achieve the volume of data presented here. So whilst this would be an idealised improvement, there must be a balance between practicality and value gained.

Furthermore, the association studies with frailty and PPI use in Chapters 3 and 4 were only carried out on a subset of the data now available within the TwinsUK cohort. These analyses would benefit from being repeated in the larger final batch of samples, which would provide more power to detect associations. Although this too would require further data collection to assay these phenotypes across this wider set of individuals.

Additional Covariates

The analyses of Chapter 6 revealed that a range of medications might influence the gut microbiota. As such, the frailty and PPI analyses should also be repeated taking into account other confounding medications. These should also be considered in future association studies. However, considering all of the medications scored in Chapter 6 would involve substantial sub-setting of samples and could produce over-fitted models. A more effective approach could be to screen future participants for medication use prior to inclusion in studies. However, especially given the number of disorders considered, it is unlikely a medication free group could be identified nor would it be representative of the natural population. Further work should instead focus on determining the minimal set of drugs that have the greatest effect on the gut microbiome, for use in recruitment screening or as covariates in analyses. The medication association results of Chapter 6 provide some information to this effect but could be improved upon by using time-matched drug use data.

A key covariate that was not collected within TwinsUK was bowel movement frequency and consistency. This has recently been shown to have a strong correlation with gut microbiota composition, having the largest effect amongst many confounders considered in the Flemish Gut study (Falony et al. 2016). This may be an important consideration, as it might influence how well stool samples represent the microbiota of areas earlier along the gastrointestinal tract. However, as with other covariates from observational studies, it will be difficult to untangle if these effects are altering the ability to accurately profile the gut microbiota or if the observed gut microbiota changes are causing the differences in transit time. Nevertheless, stool consistency should be considered as a covariate in future studies.

Limits to interrogating Causality

In the PPI analyses of Chapter 4, results were replicated in an intervention study where healthy volunteers were given PPIs, providing strong evidence for causality. However, the majority of the analyses within this thesis used data from the TwinsUK cohort or other observational studies. Where longitudinal samples were included, such as in the HMI comparisons in Chapter 6, phenotypes were only available at a single time point. By only assaying phenotypes and gut microbiota profiles at single time points it is only possible to describe associations between the two. As such, the order of cause and consequence within the observed health-microbiota associations remains unknown. Using observational data also impedes separation of the effects of correlated variables, a limitation met in the delineation of disease and drug effects in Chapter 6. Overcoming these limitations is vital to determining if gut microbiota promote or simply reflect health benefits/deficits and translation of this research to therapeutic applications.

Future work to determine causality

A combination of the approaches discussed in Section 1.2.2 would be required to determine the causal direction of the associations observed in this thesis. Especially the use of germ-free mice as a model to query the influence of gut microbiota on phenotypes of interest. This has previously been used to demonstrate causality in TwinsUK, where lean human donor samples produced delayed weight gain when transplanted into germ-free mice (Goodrich et al. 2014b). This could similarly be applied to test the hypotheses generated throughout this thesis. For example, there are several mouse models of frailty available (Seldeen, Pang, and Troen 2015). An initial comparison of the health of these models in conventional and germ-free states would provide an indication of the involvement of the gut microbiota. Further experiments to introduce the species associated with frailty in Chapter 3 would then provide evidence to their role in these effects.

Similar experiments would also be beneficial to extend the HMI analyses. For example, transplanting human microbiota from twins discordant in HMI score into germ-free mice, as in a previous study examining microbiota of twins discordant for obesity (Ridaura et al. 2013). This would enable comparison of the effects of the health-associated microbiome on murine health phenotypes such as longevity and systemic immune responses (Thevaranjan et al. 2017). This could also be repeated using immuno-compromised mice to assess interactions between the health-associated microbiome and the host immune system.

As discussed in Section 1.2.2, human FMT can also provide evidence of the causal effects of gut microbiota. Whilst, these findings are too early to alone warrant the use of FMT to improve frailty. Both the HMI and community detection approaches could be used in conjunction with existing FMT treatments/trials to assess causality. Donor samples could be profiled using both methods. The success of subsequent transplants in remedying patient

conditions would reveal if donor HMI score or the abundance of particular communities are able to predict FMT outcomes. This could be a key clinical application for the results of this thesis.

A simpler method to explore disease-microbiota dynamics in a human setting might be the use of long-term longitudinal studies. This would show if alterations in the gut microbiota proceed changes in health status or vice versa, whilst also controlling for inter-individual variation. Longitudinal analyses would be particularly useful for determining the diagnostic potential of methods such as the HMI to determine long term health outcomes. Microbiota composition has previously shown predictive power in determining response to diet and diagnostic potential in the identification of colorectal cancer (Zeevi et al. 2015; Baxter et al. 2016). However, the temporal variability of the gut microbiome and the potentially long time span over which significant changes in health status might occur, mean that this could require frequent sampling of both phenotypes and gut microbiota profiles over a long time period.

8.2.2 Wider application of results

Applicability to different demographics and body-sites

As discussed in Section 8.1.2, several novel biological phenomena were observed throughout this thesis. However, with the exception of analyses using the Israeli data in Chapter 7, these were exclusively based on data from adult Western populations. Given the known differences by geography and age in the gut microbiome, this may limit the application of these results to other demographics. The presented analyses would be improved by inclusion of replication datasets from a more diverse range of populations. Future work should be carried out to this effect.

The described studies largely considered data from faecal samples or swabs. The microbiome varies in composition along the length of the GI tract (Stearns et al. 2011). Therefore the described phenomena may be specific to colonic microbiota. Further studies using biopsies and/or swabs to sample other sites would be required to determine the locality of the observed effects. More research is also warranted to see if consistent health associations and community structures can be identified in microbiota at other non-GI body sites.

Technical limits to clinical application

The majority of phenotypes queried in these studies were either directly health related, such as disease states and medication use, or health associated, such as age and BMI. However, the results are insufficient for direct clinical application. Not only are further experiments required to understand the mechanisms underlying the associations (as discussed previously)

the experimental and analytical pipelines used for 16S rRNA gene sequencing approaches are currently limited to academia. Application of these findings to a clinical setting will require further developments of rapid, cheap, reproducible, and simple approaches to 16S rRNA gene amplicon-based microbiota profiling. Alternatively, development of targeted approaches, such as targeted PCR or micro-array assays (Tottey et al. 2013) to selectively quantify markers, such as those in the HMI, could negate the requirement for wide-scale 16S rRNA gene sequencing.

Limits to wider adoption of methods

Ideally the methods described here will be adopted by the microbiome research community. However, there are a number of factors limiting wider application of these techniques.

Realistically, the wide-scale adoption of heritability as a measure of OTU clustering quality is unlikely. It requires host genomic or familial data to estimate heritability and, unlike the stability and accuracy measures used in other clustering comparisons (He et al. 2015; Westcott and Schloss 2015), there is no direct interpretation of what is measured. Although the results were in concordance with those of existing studies (Westcott and Schloss 2015) and, at least within the TwinsUK data, it proved an apt measure to ensure optimal OTUs were generated for use in Chapters 6 and 7.

The co-occurrence network method described in Chapter 7 may also be too complex for wider use. It uses methods that are novel to microbiome research and includes several complex parameter selection steps. This could restrict its uptake outside the community of researchers already investigating co-occurrence networks in the gut microbiome. Existing bioinformatics packages that have been widely adopted by the community, such as QIIME and Mothur (Navas-Molina et al. 2013; Schloss et al. 2009), have done so by presenting simple and accessible interfaces. Future work to promote wider use of the community detection approach in Chapter 7 should focus on simplifying the technique and consolidating the stages into a single accessible workflow. Simplification could be achieved by using one co-occurrence metric, such as CoNet, and further testing to identify if the selected parameters are more widely applicable, negating the requirement for complex parametrisation steps.

The HMI could also benefit from the creation of a simple and accessible script to generate the index from family level summaries. Whilst this could be easily achieved, QIIME already has functionality to generate the MDI in the script '*compute_taxonomy_ratios*'. This simply requires an alternate input list of familial definitions to generate the HMI (Navas-Molina et al. 2013). The HMI and health-associated microbiome definition are the most accessible methods presented in this thesis.

A further limit to adoption of these methods within the microbiome research community is the increased use of metagenomics profiling in place of 16S rRNA gene sequencing ap-

proaches. As analytical pipelines become more accessible and efficient and costs lower, there is further potential for metagenomics to be used in place of amplicon-based profiling. However, the techniques described in this thesis can, and should, be applied to metagenomic datasets. I discuss this is more detail in Section 8.2.3.

8.2.3 Alternate approaches to microbiome profiling

Alternative 16S rRNA gene sequencing approaches

Throughout this thesis I have tried to use optimal methods to analyse data. This was achieved by either comparing available options directly, as in the OTU algorithm comparisons in Chapter 5, or referencing existing benchmarking studies, such as the Weiss *et al.* study used in Chapter 7 (Weiss et al. 2016). However, these are not comprehensive comparisons. For instance in Chapter 5, as the TwinsUK data were so large, I did not consider the hierarchical clustering approaches used by the popular Mothur platform (Schloss et al. 2009). There have also been new approaches developed since these comparisons and non-clustering approaches such as DADA2 have gained wider adoption (Callahan et al. 2016). Similarly, the Weiss *et al.* comparison did not consider several existing approaches to quantify microbial co-occurrence (Fang et al. 2015; Kurtz et al. 2015). To improve upon these limitations I would re-approach these comparisons using a smaller standardised dataset. This would increase the range of approaches that could be considered and facilitate future inclusion of new techniques. For the purposes of this thesis I was only interested in approaches I could apply to TwinsUK for use in other methods. To this end the comparisons were sufficient.

I have also developed methods within this thesis that could be applied to one another. For instance, identifying communities of microbiota associated with the HMI. Whilst this would be a trivial analysis, it would be largely uninformative as communities were largely homogeneous at the family level. As such, I would expect them to associate with the HMI based on the health/disorder classification of their dominant family. However, the contrasting analysis would be particularly interesting. In the future, the multi-disease analysis in Chapter 6 should be repeated using communities (as described in Chapter 7) in place of the marker traits. This would enable identification of microbial communities that are consistently health or disorder associated. These could be key components in the relationship between the gut microbiota and human health.

Application to metagenomics data

I have shown that it is possible to improve the analysis and interpretation of 16S rRNA gene sequencing data sets. However as discussed in Section 1.3.3, all marker gene techniques

are inherently limited in their in ability to resolve fine-level taxonomic differences or functional aspects of the gut microbiome. Whilst the methods in this thesis were developed for 16S rRNA gene amplicon profiles, they could easily be adapted to profiles from shotgun metagenomic sequencing. In particular, future work should focus on extending the analyses described in Chapters 6 and 7, to apply the multi-disorder and community detection approaches to metagenomic data.

The analysis of multiple diseases in Chapter 6 could be extended by using shotgun metagenomics to profile gut microbiota. I do not see value in using higher taxonomic resolutions to define the health-associated microbiome. I believe the strength of the definition comes from its broadness. There is also higher within-individual variability at finer taxonomic levels (Consortium 2012). However, there is a known disparity in the taxonomic estimates provided by metagenomic and 16S rRNA gene derived microbiota profiles (Jovel et al. 2016). So this analysis would be required to determine the effects this has on the familial health-associated microbiome definition and quantification of the HMI from metagenomic data.

One area I would be interested to explore further, is the use of metagenomic profiles to probe gut microbiome functionality across multiple diseases. Microbiome function is more highly conserved than taxonomic abundances (Consortium 2012). Future analyses should focus on the identification of microbiome functions that are consistently health or disorder associated. This would be possible using similar analyses as with the marker taxa in Chapter 6, but using the abundance of metagenomic pathways. This could help uncover the microbial mechanisms underlying the consistent gut microbiota associations across multiple diseases.

In contrast to the HMI study, I think that the community detection approach from Chapter 7 would benefit from the high resolution taxonomic profiles provided by metagenomic sequencing. The community units produced in Chapter 7 are able to reduce the dimensionality of 16S rRNA gene sequencing data and reflect the biological interactions observed across samples. However, they provide little information regarding their constituent taxa, the mechanisms of interaction, or the communities' functionalities. These could all be addressed using shotgun metagenomics. Precise identification of species would also further aid applications of this approach to the design of synthetic communities as discussed in Section 8.1.1.

Any future approaches using metagenomics would also benefit from concurrent use of metabolomics or metatranscriptomics. Metagenomic sequencing can only infer potential microbial function, where as these approaches reflect active functionality. This would allow further investigation of the mechanisms underlying the effects described throughout this thesis.

8.3 Conclusion

In this thesis I have presented novel approaches to address the limitations of dimensionality, reproducibility, and biological relevance associated with gut microbiota profiles derived from 16S rRNA gene sequencing. These include the development of the HMI, an index representing both microbiota composition and host health, and a method to identify stable microbial communities within microbiota. In this process I have also identified several novel biological phenomena including: associations of host frailty and PPI use with the gut microbiota, the consistency of microbiota associations across multiple diseases, and the stability of gut microbiota community structures across populations. These findings advance our understanding of the gut microbiota's role in human health and provide novel tools and insights facilitating further research in this exciting, fast moving, and important area.

References

Acharya, Nandini et al. (2017). "Endocannabinoid system acts as a regulator of immune homeostasis in the gut". In: *Proceedings of the National Academy of Sciences*.

Aitchison, John (1982). "The statistical analysis of compositional data". In: *Journal of the Royal Statistical Society. Series B* 44, p. 139.

Albert, Reka (2005). "Scale-free networks in cell biology". In: *Journal of Cell Science* 118.21, p. 4947.

Alexander, James L et al. (2017). "Gut microbiota modulation of chemotherapy efficacy and toxicity". In: *Nature Reviews Gastroenterology and Hepatology* 14.6, p. 356.

Ananthakrishnan, Ashwin N (2015). "Epidemiology and risk factors for IBD". In: *Nature Reviews Gastroenterology & Hepatology* 12.4, p. 205.

Arumugam, Manimozhiyan et al. (2011). "Enterotypes of the human gut microbiome". In: *Nature* 473, p. 174.

Aßhauer, Kathrin P. et al. (2015). "Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data". In: *Bioinformatics* 31.17, p. 2882.

Bäckhed, Fredrik et al. (2007). "Mechanisms underlying the resistance to diet-induced obesity in germ-free mice". In: *Proceedings of the National Academy of Sciences* 104.3, p. 979.

Bäckhed, Fredrik et al. (2015). "Dynamics and stabilization of the human gut microbiome during the first year of life". In: *Cell Host & Microbe* 17.5, p. 690.

Baldassano, Steven N and Danielle S Bassett (2016). "Topological distortion and reorganized modular structure of gut microbial co-occurrence networks in inflammatory bowel disease". In: *Scientific Reports* 6, p. 26087.

Bastian, Mathieu, Sebastien Heymann, Mathieu Jacomy, et al. (2009). "Gephi: an open source software for exploring and manipulating networks." In: *International AAAI Conference on Web and Social Media* 8, p. 361.

Baxter, Nielson T et al. (2016). "Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions". In: *Genome Medicine* 8.1, p. 37.

Beserra, Bruna TS et al. (2015). "A systematic review and meta-analysis of the prebiotics and synbiotics effects on glycaemia, insulin concentrations and lipid parameters in adult patients with overweight or obesity". In: *Clinical Nutrition* 34.5, p. 845.

Biagi, Elena et al. (2010). "Through ageing, and beyond: gut microbiota and inflammatory status in seniors and centenarians". In: *PLoS One* 5.5, e10667.

Blekhman, Ran et al. (2015). "Host genetic variation impacts microbiome composition across human body sites". In: *Genome Biology* 16.1, p. 191.

Blondel, Vincent D et al. (2008). "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10, p. 10008.

Boker, Steven et al. (2011). "OpenMx: an open source extended structural equation modeling framework". In: *Psychometrika* 76.2, p. 306.

Bonder, Marc Jan et al. (2016). "The effect of host genetics on the gut microbiome". In: *Nature Genetics* 48.11, p. 1407.

Boulangé, Claire L et al. (2016). "Impact of the gut microbiota on inflammation, obesity, and metabolic disease". In: *Genome Medicine* 8.1, p. 42.

Bowles, Nicole P et al. (2015). "A peripheral endocannabinoid mechanism contributes to glucocorticoid-mediated metabolic syndrome". In: *Proceedings of the National Academy of Sciences* 112.1, p. 285.

Brandt, Lawrence J et al. (2012). "Long-term follow-up of colonoscopic fecal microbiota transplant for recurrent Clostridium difficile infection". In: *The American Journal of Gastroenterology* 107.7, p. 1079.

Brosius, Jurgen et al. (1978). "Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli". In: *Proceedings of the National Academy of Sciences* 75.10, p. 4801.

Browne, Hilary P et al. (2016). "Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation." In: *Nature* 533.7604, p. 543.

Buffie, Charlie G et al. (2015). "Precision microbiome restoration of bile acid-mediated resistance to Clostridium difficile". In: *Nature* 517, p. 205.

Callahan, Benjamin J et al. (2016). "DADA2: high-resolution sample inference from Illumina amplicon data". In: *Nature Methods* 13.7, p. 581.

Cani, Patrice D et al. (2007). "Metabolic endotoxemia initiates obesity and insulin resistance". In: *Diabetes* 56.7, p. 1761.

Cani, Patrice D et al. (2008). "Changes in gut microbiota control metabolic endotoxemia-induced inflammation in high-fat diet–induced obesity and diabetes in mice". In: *Diabetes* 57.6, p. 1470.

Caporaso, J Gregory et al. (2011a). "Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample". In: *Proceedings of the National Academy of Sciences* 108, p. 4516.

Caporaso, J. Gregory et al. (2011b). "Moving pictures of the human microbiome". In: *Genome Biology* 12.5, R50.

Chaffron, Samuel et al. (2010). "A global network of coexisting microbes from environmental and whole-genome sequence data". In: *Genome Research* 20.7, p. 947.

Chao, Anne (1984). "Nonparametric estimation of the number of classes in a population". In: *Scandinavian Journal of Statistics*, p. 265.

Chen, Mingming, Konstantin Kuzmin, and Boleslaw K Szymanski (2014). "Community detection via maximization of modularity and its variants". In: *IEEE Transactions on Computational Social Systems* 1.1, p. 46.

Chen, Wei et al. (2013). "A comparison of methods for clustering 16S rRNA sequences into OTUs". In: *PLoS One* 8.8, e70837.

Claesson, Marcus J et al. (2012). "Gut microbiota composition correlates with diet and health in the elderly". In: *Nature* 492, p. 178.

Cole, James R et al. (2008). "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis". In: *Nucleic Acids Research* 37.S1, p. D141.

Consortium, The Human Microbiome Project (2012). "Structure, function and diversity of the healthy human microbiome". In: *Nature* 486, p. 207.

Corrêa-Oliveira, Renan et al. (2016). "Regulation of immune cell function by short-chain fatty acids". In: *Clinical & Translational Immunology* 5.4.

Costello, Elizabeth K et al. (2009). "Bacterial community variation in human body habitats across space and time". In: *Science* 326.5960, pp. 1694–1697.

Crost, Emmanuelle H et al. (2013). "Utilisation of mucin glycans by the human gut symbiont Ruminococcus gnavus is strain-dependent". In: *PloS one* 8.10, e76341.

Cruaud, Perrine et al. (2014). "Influence of DNA extraction method, 16S rRNA targeted hypervariable regions, and sample origin on microbial diversity detected by 454 pyrosequencing in marine chemosynthetic ecosystems". In: *Applied and Environmental Microbiology* 80.15, p. 4626.

Csardi, Gabor and Tamas Nepusz (2006). "The igraph software package for complex network research". In: *InterJournal, Complex Systems* 1695.5, p. 1.

D'Argenio, Giuseppe et al. (2007). "Overactivity of the intestinal endocannabinoid system in celiac disease and in methotrexate-treated rats". In: *Journal of Molecular Medicine* 85.5, p. 523.

David, Lawrence A et al. (2014). "Diet rapidly and reproducibly alters the human gut microbiome". In: *Nature* 505, p. 559.

De Roy, Karen et al. (2014). "Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities". In: *Environmental Microbiology* 16.6, p. 1472.

Deng, Ye et al. (2012). "Molecular ecological network analyses". In: *BMC Bioinformatics* 13.1, p. 113.

Desai, Mahesh S et al. (2016). "A dietary fiber-deprived gut microbiota degrades the colonic mucus barrier and enhances pathogen susceptibility". In: *Cell* 167.5, p. 1339.

DeSantis, Todd Z et al. (2006). "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB". In: *Applied and Environmental Microbiology* 72.7, p. 5069.

Di Marzo, V and AA Izzo (2006). "Endocannabinoid overactivity and intestinal inflammation". In: *Gut* 55.10, p. 1373.

Donia, Mohamed S et al. (2014). "A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics". In: *Cell* 158.6, p. 1402.

Donohoe, Dallas R et al. (2012). "The Warburg effect dictates the mechanism of butyrate-mediated histone acetylation and cell proliferation". In: *Molecular cell* 48.4, pp. 612–626.

Duran-Pinedo, Ana E et al. (2011). "Correlation network analysis applied to complex biofilm communities". In: *PLoS One* 6.12, e28438.

Edgar, Robert C (2010). "Search and clustering orders of magnitude faster than BLAST". In: *Bioinformatics* 26.19, p. 2460.

Edgar, Robert C et al. (2011). "UCHIME improves sensitivity and speed of chimera detection". In: *Bioinformatics* 27.16, p. 2194.

Ehrlich, Garth D, N Luisa Hiller, and Fen Ze Hu (2008). "What makes pathogens pathogenic". In: *Genome Biology* 9.6, p. 225.

Eren, A Murat et al. (2015). "Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences". In: *The ISME Journal* 9.4, p. 968.

Faith, Daniel P and Andrew M Baker (2006). "Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges". In: *Evolutionary Bioinformatics* 2, p. 121.

Falony, Gwen et al. (2016). "Population-level analysis of gut microbiome variation". In: *Science* 352.6285, p. 560.

Fang, Huaying et al. (2015). "CCLasso: correlation inference for compositional data through Lasso". In: *Bioinformatics* 31.19, p. 3172.

Faust, Karoline and Jeroen Raes (2012). "Microbial interactions: from networks to models". In: *Nature Reviews Microbiology* 10.8, p. 538.

Faust, Karoline et al. (2012). "Microbial co-occurrence relationships in the human microbiome". In: *PLoS Computational Biology* 8.7, e1002606.

Faust, Karoline et al. (2015). "Metagenomics meets time series analysis: unraveling microbial community dynamics". In: *Current Opinion in Microbiology* 25, p. 56.

Ford, Alexander C et al. (2014). "Efficacy of prebiotics, probiotics, and synbiotics in irritable bowel syndrome and chronic idiopathic constipation: systematic review and meta-analysis". In: *The American journal of gastroenterology* 109.10, p. 1547.

Forslund, Kristoffer et al. (2015). "Disentangling the effects of type 2 diabetes and metformin on the human gut microbiota". In: *Nature* 528, p. 262.

Frank, Daniel N et al. (2011). "Disease phenotype and genotype are associated with shifts in intestinal-associated microbiota in inflammatory bowel diseases". In: *Inflammatory Bowel Diseases* 17.1, p. 179.

Freedberg, Daniel E et al. (2015). "Proton pump inhibitors alter specific taxa in the human gastrointestinal microbiome: a crossover trial". In: *Gastroenterology* 149.4, p. 883.

Frias-Lopez, Jorge et al. (2008). "Microbial community gene expression in ocean surface waters". In: *Proceedings of the National Academy of Sciences* 105.10, p. 3805.

Friedman, Jonathan and Eric J Alm (2012). "Inferring correlation networks from genomic survey data". In: *PLoS Computational Biology* 8.9, e1002687.

Friedman, Jonathan, Logan M. Higgins, and Jeff Gore (2017). "Community structure follows simple assembly rules in microbial microcosms". In: *Nature Ecology & Evolution* 1.5.

Fu, Jingyaun et al. (2015). "The gut microbiome contributes to a substantial proportion of the variation in blood lipids". In: *Circulation Research*, p. 115.

Gao, Zhanguo et al. (2009). "Butyrate improves insulin sensitivity and increases energy expenditure in mice". In: *Diabetes* 58.7, p. 1509.

Gevers, Dirk et al. (2014). "The treatment-naive microbiome in new-onset Crohn's disease". In: *Cell Host & Microbe* 15.3, p. 382.

Gilbert, Jack A, Janet K Jansson, and Rob Knight (2014). "The Earth Microbiome project: successes and aspirations". In: *BMC Biology* 12.1, p. 69.

Gloor, Gregory B et al. (2016). "It's all relative: analyzing microbiome data as compositions". In: *Annals of Epidemiology* 26.5, p. 322.

Gohl, Daryl M et al. (2016). "Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies". In: *Nature Biotechnology* 201, p. 6.

Goodrich, Julia K et al. (2014a). "Conducting a microbiome study". In: *Cell* 158.2, p. 250.

Goodrich, Julia K et al. (2014b). "Human genetics shape the gut microbiome". In: *Cell* 159.4, p. 789.

Goodwin, Sara, John D McPherson, and W Richard McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies". In: *Nature Reviews Genetics* 17.6, p. 333.

Grehan, Martin J et al. (2010). "Durable alteration of the colonic microbiota by the administration of donor fecal flora". In: *Journal of Clinical Gastroenterology* 44.8, p. 551.

Gupta, Vinod K, Sandip Paul, and Chitra Dutta (2017). "Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity". In: *Frontiers in Microbiology* 8, p. 1162.

Haas, Brian J et al. (2011). "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons". In: *Genome Research* 21.3, p. 494.

Hakansson, Asa and Goran Molin (2011). "Gut microbiota and inflammation". In: *Nutrients* 3.6, p. 637.

Halfvarson, Jonas et al. (2017). "Dynamics of the human gut microbiome in inflammatory bowel disease". In: *Nature Microbiology* 2, p. 17004.

Harrell Jr, Frank E, Charles Dupont, et al. (2008). "Hmisc: harrell miscellaneous". In: *R package version* 3.2.

He, Yan et al. (2015). "Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity". In: *Microbiome* 3.1, p. 20.

Hjorth, MF et al. (2017). "Pre-treatment microbial Prevotella-to-Bacteroides ratio, determines body fat loss success during a 6-month randomized controlled diet intervention." In: *International Journal of Obesity*.

Hsiao, Elaine Y et al. (2013). "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders". In: *Cell* 155.7, p. 1451.

Hsu, David et al. (2010). "Toll-like receptor 4 differentially regulates epidermal growth factor-related growth factors in response to intestinal mucosal injury". In: *Laboratory Investigation* 90.9, p. 1295.

Hubbard, Ruth E and Ken W Woodhouse (2010). "Frailty, inflammation and the elderly". In: *Biogerontology* 11.5, p. 635.

Huddleston, Jennifer R (2014). "Horizontal gene transfer in the human gastrointestinal tract: potential spread of antibiotic resistance genes". In: *Infection and Drug Resistance* 7, p. 167.

Huse, Susan M et al. (2010). "Ironing out the wrinkles in the rare biosphere through improved OTU clustering". In: *Environmental Microbiology* 12.7, p. 1889.

Imhann, Floris et al. (2016). "Proton pump inhibitors affect the gut microbiome". In: *Gut* 65.5, p. 740.

Islam, KBM Saiful et al. (2011). "Bile acid is a host factor that regulates the composition of the cecal microbiota in rats". In: *Gastroenterology* 141.5, p. 1773.

Jakobsson, Hedvig E et al. (2010). "Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome". In: *PLoS One* 5.3, e9836.

Jeffery, Ian B, Denise B Lynch, and Paul W O'Toole (2016). "Composition and temporal stability of the gut microbiota in older persons". In: *The ISME Journal* 10.1, p. 170.

Jeffery, Ian B et al. (2012). "Categorization of the gut microbiota: enterotypes or gradients?" In: *Nature Reviews Microbiology* 10.9, p. 591.

Jeong, H et al. (2000). "The large-scale organization of metabolic networks". In: *Nature* 407.6804, p. 651.

Jernberg, Cecilia et al. (2007). "Long-term ecological impacts of antibiotic administration on the human intestinal microbiota". In: *The ISME Journal* 1.1, p. 56.

Jiménez, Esther et al. (2008). "Is meconium from healthy newborns actually sterile?" In: *Research in Microbiology* 159.3, p. 187.

Jones, Susan (2013). *Trends in Microbiome Research*.

Jovel, Juan et al. (2016). "Characterization of the gut microbiome using 16S or shotgun metagenomics". In: *Frontiers in Microbiology* 7.

Kadouri, Daniel E et al. (2013). "Predatory bacteria: a potential ally against multidrug-resistant Gram-negative pathogens". In: *PLoS One* 8.5, e63397.

Kakiyama, Genta et al. (2013). "Modulation of the fecal bile acid profile by gut microbiota in cirrhosis". In: *Journal of hepatology* 58.5, p. 949.

Kamada, Nobuhiko et al. (2012). "Regulated virulence controls the ability of a pathogen to compete with the gut microbiota". In: *Science* 336.6086, p. 1325.

Kassinen, Anna et al. (2007). "The fecal microbiota of irritable bowel syndrome patients differs significantly from that of healthy subjects". In: *Gastroenterology* 133.1, p. 24.

Kelsic, Eric D et al. (2015). "Counteraction of antibiotic production and degradation stabilizes microbial communities". In: *Nature* 521.7553, p. 516.

Kerr, Benjamin et al. (2002). "Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors". In: *Nature* 418, p. 171.

Kim, Mincheol et al. (2014). "Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes". In: *International Journal of Systematic and Eevolutionary Microbiology* 64.2, p. 346.

Kim, Minseok, Mark Morrison, and Zhongtang Yu (2011). "Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes". In: *Journal of Microbiological Methods* 84.1, p. 81.

Knight, Rob, Jeff Leach, and The American Gut Consortium. *The American Gut Project*. URL: `ftp://ftp.microbio.me/AmericanGut`.

Koenig, Jeremy E et al. (2011). "Succession of microbial consortia in the developing infant gut microbiome". In: *Proceedings of the National Academy of Sciences* 108.Supplement 1, p. 4578.

Kommineni, Sushma et al. (2015). "Bacteriocin production augments niche competition by enterococci in the mammalian gastrointestinal tract". In: *Nature* 526.7575, p. 719.

Konopka, Allan, Stephen Lindemann, and Jim Fredrickson (2015). "Dynamics in microbial communities: unraveling mechanisms to identify principles". In: *The ISME Journal* 9.7, p. 1488.

Konstantinidis, Konstantinos T and James M Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes". In: *Proceedings of the National Academy of Sciences* 102.7, p. 2567.

Kopylova, Evguenia et al. (2016). "Open-source sequence clustering methods improve the state of the art". In: *mSystems* 1.1, e00003.

Koslowski, Maureen J et al. (2010). "Innate antimicrobial host defense in small intestinal Crohn's disease". In: *International Journal of Medical Microbiology* 300.1, p. 34.

Kovatcheva-Datchary, Petia et al. (2015). "Dietary fiber-induced improvement in glucose metabolism is associated with increased abundance of Prevotella". In: *Cell Metabolism* 22.6, p. 971.

Kultima, Jens Roat et al. (2012). "MOCAT: a metagenomics assembly and gene prediction toolkit". In: *PLoS One* 7.10, e47656.

Kurtz, Zachary D et al. (2015). "Sparse and compositionally robust inference of microbial ecological networks". In: *PLoS Computational Biology* 11.5, e1004226.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1, p. 559.

Langille, Morgan GI et al. (2013). "Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences". In: *Nature Biotechnology* 31.9, p. 814.

Larsen, Nadja et al. (2010). "Gut microbiota in human adults with type 2 diabetes differs from non-diabetic adults". In: *PLoS One* 5.2, e9085.

Larsen, Peter E, Dawn Field, and Jack A Gilbert (2012). "Predicting bacterial community assemblages using an artificial neural network approach". In: *Nature Methods* 9.6, p. 621.

Lederberg, Joshua and Alexa T McCray (2001). "Ome SweetOmics–A Genealogical Treasury of Words". In: *The Scientist* 15.7.

Leffler, Daniel A and J Thomas Lamont (2015). "Clostridium difficile infection". In: *New England Journal of Medicine* 372.16, p. 1539.

Levy, Roie and Elhanan Borenstein (2013). "Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules". In: *Proceedings of the National Academy of Sciences* 110.31, p. 12804.

— (2014). "Metagenomic systems biology and metabolic modeling of the human microbiome: From species composition to community assembly rules". In: *Gut Microbes* 5.2, p. 265.

Ley, Ruth E et al. (2005). "Obesity alters gut microbial ecology". In: *Proceedings of the National Academy of Sciences* 102.31, p. 11070.

Ley, Ruth E et al. (2006). "Human gut microbes associated with obesity". In: *Nature* 444.7122, p. 1022.

Li, Jia V et al. (2011). "Metabolic surgery profoundly influences gut microbial–host metabolic cross-talk". In: *Gut* 60.9, p. 1214.

Li, Junhua et al. (2014). "An integrated catalog of reference genes in the human gut microbiome". In: *Nature Biotechnology* 32.8, p. 834.

Li, Weizhong and Adam Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences". In: *Bioinformatics* 22.13, p. 1658.

Lim, Mi Young et al. (2016). "The effect of heritability and host genetics on the gut microbiota and metabolic syndrome". In: *Gut* 66.6, p. 1031.

Lin, Ching Yu et al. (2007). "Evaluation of metabolite extraction strategies from tissue samples using NMR metabolomics". In: *Metabolomics* 3.1, p. 55.

Lindemann, Stephen R et al. (2016). "Engineering microbial consortia for controllable outputs". In: *The ISME Journal* 10.9, p. 2077.

Lloyd-Price, Jason, Galeb Abu-Ali, and Curtis Huttenhower (2016). "The healthy human microbiome". In: *Genome Medicine* 8.1, p. 51.

Louis, Petra and Harry J Flint (2009). "Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine". In: *FEMS Microbiology Letters* 294.1, p. 1.

Louis, Petra, Georgina L Hold, and Harry J Flint (2014). "The gut microbiota, bacterial metabolites and colorectal cancer". In: *Nature Reviews Microbiology* 12.10, p. 661.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550.

Lozupone, Catherine and Rob Knight (2005). "UniFrac: a new phylogenetic method for comparing microbial communities". In: *Applied and Environmental Microbiology* 71.12, p. 8228.

Lozupone, Catherine et al. (2012a). "Identifying genomic and metabolic features that can underlie early successional and opportunistic lifestyles of human gut symbionts". In: *Genome Research* 22.10, p. 1974.

Lozupone, Catherine A et al. (2012b). "Diversity, stability and resilience of the human gut microbiota". In: *Nature* 489.7415, p. 220.

Lukashin, Alexander V and Mark Borodovsky (1998). "GeneMark. hmm: new solutions for gene finding". In: *Nucleic Acids Research* 26.4, p. 1107.

Luo, Feng et al. (2006). "Application of random matrix theory to biological networks". In: *Physics Letters A* 357.6, p. 420.

Lynch, Susan V and Oluf Pedersen (2016). "The human intestinal microbiome in health and disease". In: *New England Journal of Medicine* 375.24, p. 2369.

MacKenzie, Donald A et al. (2010). "Strain-specific diversity of mucus-binding proteins in the adhesion and aggregation properties of Lactobacillus reuteri". In: *Microbiology* 156.11, p. 3368.

Mahé, Frédéric et al. (2014). "Swarm: robust and fast clustering method for amplicon-based studies". In: *PeerJ* 2, e593.

Maloy, Kevin J and Fiona Powrie (2011). "Intestinal homeostasis and its breakdown in inflammatory bowel disease". In: *Nature* 474.7351, p. 298.

Manichanh, Chaysavanh et al. (2006). "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach". In: *Gut* 55.2, p. 205.

Marchesi, Julian R and Jacques Ravel (2015). "The vocabulary of microbiome research: a proposal". In: *Microbiome* 3.1, p. 31.

Mardinoglu, Adil et al. (2015). "The gut microbiota modulates host amino acid and glutathione metabolism in mice". In: *Molecular Systems Biology* 11.10, p. 834.

Maslowski, Kendle M et al. (2009). "Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43". In: *Nature* 461.7268, p. 1282.

Mazmanian, Sarkis K, June L Round, and Dennis L Kasper (2008). "A microbial symbiosis factor prevents intestinal inflammatory disease". In: *Nature* 453.7195, p. 620.

Mazmanian, Sarkis K et al. (2005). "An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system". In: *Cell* 122.1, p. 107.

McGeachie, Michael J, Hsun-Hsien Chang, and Scott T Weiss (2014). "CGBayesNets: conditional Gaussian Bayesian network learning and inference with mixed discrete and continuous data". In: *PLoS Computational Biology* 10.6, e1003676.

McGeachie, Michael J et al. (2016). "Longitudinal prediction of the infant gut microbiome with dynamic bayesian networks". In: *Scientific Reports* 6.

McHardy, Ian H et al. (2013). "Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships". In: *Microbiome* 1.1, p. 17.

McMurdie, Paul J and Susan Holmes (2014). "Waste not, want not: why rarefying microbiome data is inadmissible". In: *PLoS Computational Biology* 10.4, e1003531.

Mercier, Céline et al. (2013). "SUMATRA and SUMACLUST: fast and exact comparison and clustering of sequences". In: *Programs and Abstracts of the SeqBio 2013 workshop. Abstract*, p. 27.

Michail, Sonia et al. (2012). "Alterations in the gut microbiome of children with severe ulcerative colitis". In: *Inflammatory Bowel Diseases* 18.10, p. 1799.

Mizoguchi, Atsushi (2012). "Animal models of inflammatory bowel disease." In: *Progress in Molecular Biology and Translational Science* 105, p. 263.

Moayyeri, Alireza et al. (2012). "Cohort profile: TwinsUK and healthy ageing twin study". In: *International Journal of Epidemiology* 42.1, p. 76.

Moles, Laura et al. (2013). "Bacterial diversity in meconium of preterm neonates and evolution of their fecal microbiota during the first month of life". In: *PloS One* 8.6, e66986.

Morris, Brandon EL et al. (2013). "Microbial syntrophy: interaction for the common good". In: *FEMS Microbiology Reviews* 37.3, p. 384.

Mougi, Akihiko (2016). "The roles of amensalistic and commensalistic interactions in large ecological network stability". In: *Scientific Reports* 6.

Muccioli, Giulio G et al. (2010). "The endocannabinoid system links gut microbiota to adipogenesis". In: *Molecular Systems Biology* 6.1, p. 392.

Mucha, Peter J et al. (2010). "Community structure in time-dependent, multiscale, and multiplex networks". In: *Science* 328.5980, p. 876.

Mudaliar, Sunder et al. (2013). "Efficacy and safety of the farnesoid X receptor agonist obeticholic acid in patients with type 2 diabetes and nonalcoholic fatty liver disease". In: *Gastroenterology* 145.3, p. 574.

Mysara, Mohamed et al. (2017). "Reconciliation between operational taxonomic units and species boundaries". In: *FEMS Microbiology Ecology* 93.4.

Navas-Molina, José A et al. (2013). "Advancing our understanding of the human microbiome using QIIME". In: *Methods in Enzymology* 531, p. 371.

Neis, Evelien PJG, Cornelis HC Dejong, and Sander S Rensen (2015). "The role of microbial amino acid metabolism in host metabolism". In: *Nutrients* 7.4, p. 2930.

Newman, Mark EJ (2004). "Detecting community structure in networks". In: *The European Physical Journal B-Condensed Matter and Complex Systems* 38.2, p. 321.

— (2006). "Modularity and community structure in networks". In: *Proceedings of the National Academy of Sciences* 103.23, p. 8577.

Newman, Mark EJ and Michelle Girvan (2004). "Finding and evaluating community structure in networks". In: *Phys. Rev. E* 69 (2), p. 026113.

Ng, Siew Chien et al. (2013). "Effect of probiotic bacteria on the intestinal microbiota in irritable bowel syndrome". In: *Journal of Gastroenterology and Hepatology* 28.10, p. 1624.

Nguyen, Nam-Phuong et al. (2016). "A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity". In: *Biofilms and Microbiomes* 2, p. 16004.

Odamaki, Toshitaka et al. (2016). "Age-related changes in gut microbiota composition from newborn to centenarian: a cross-sectional study". In: *BMC Microbiology* 16.1, p. 90.

Oksanen, Jari et al. (2016). *vegan: Community Ecology Package*. R package version 2.4-1. URL: https://CRAN.R-project.org/package=vegan.

O'Neill, Ian, Zoe Schofield, and Lindsay J Hall (2017). "Exploring the role of the microbiota member Bifidobacterium in modulating immune-linked diseases". In: *Emerging Topics in Life Sciences* 1.4, p. 333.

Paliy, Oleg et al. (2009). "High-throughput quantitative analysis of the human intestinal microbiota with a phylogenetic microarray". In: *Applied and Environmental Microbiology* 75.11, p. 3572.

Park, Jung-ha et al. (2016). "Promotion of intestinal epithelial cell turnover by commensal bacteria: role of short-chain fatty acids". In: *PloS one* 11.5, e0156334.

Pimentel, Mark et al. (2012). "Methanogens in human health and disease". In: *The American Journal of Gastroenterology* 1.1, p. 28.

Pozuelo, Marta et al. (2015). "Reduction of butyrate-and methane-producing microorganisms in patients with irritable bowel syndrome". In: *Scientific Reports* 5.

Prakash, Tulika and Todd D Taylor (2012). "Functional assignment of metagenomic data: challenges and applications". In: *Briefings in Bioinformatics* 13.6, p. 711.

Qin, Junjie et al. (2010). "A human gut microbial gene catalog established by metagenomic sequencing". In: *Nature* 464, p. 59.

Qin, Junjie et al. (2012). "A metagenome-wide association study of gut microbiota in type 2 diabetes". In: *Nature* 490, p. 55.

Qin, Nan et al. (2014). "Alterations of the human gut microbiome in liver cirrhosis". In: *Nature* 513, p. 59.

Qiu, Xiaoyun et al. (2001). "Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning". In: *Applied and Environmental microbiology* 67.2, p. 880.

Quast, Christian et al. (2012). "The SILVA ribosomal RNA gene database project: improved data processing and web-based tools". In: *Nucleic Acids Research* 41.D1, p. D590.

Rajilić-Stojanović, Mirjana et al. (2011). "Global and deep molecular analysis of microbiota signatures in fecal samples from patients with irritable bowel syndrome". In: *Gastroenterology* 141.5, pp. 1792–1801.

Rampelli, Simone et al. (2015). "Metagenome sequencing of the Hadza hunter-gatherer gut microbiota". In: *Current Biology* 25.13, p. 1682.

Rausch, Philipp et al. (2011). "Colonic mucosa-associated microbiota is influenced by an interaction of Crohn disease and FUT2 (Secretor) genotype". In: *Proceedings of the National Academy of Sciences* 108.47, p. 19030.

Rea, Mary C et al. (2010). "Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against Clostridium difficile". In: *Proceedings of the National Academy of Sciences* 107.20, p. 9352.

Reshef, David N et al. (2011). "Detecting novel associations in large data sets". In: *Science* 334.6062, p. 1518.

Ridaura, Vanessa K et al. (2013). "Gut microbiota from twins discordant for obesity modulate metabolism in mice". In: *Science* 341.6150, p. 1241214.

Rideout, Jai Ram et al. (2014). "Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences". In: *PeerJ* 2, e545.

Ridlon, Jason M et al. (2014). "Bile acids and the gut microbiome". In: *Current Opinion in Gastroenterology* 30.3, p. 332.

Rognes, Torbjørn et al. (2016). "VSEARCH: a versatile open source tool for metagenomics". In: *PeerJ* 4, e2584.

Round, June L and Sarkis K Mazmanian (2009). "The gut microbiome shapes intestinal immune responses during health and disease". In: *Nature Reviews Immunology* 9.5, p. 313.

Ruan, Quansong et al. (2006). "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors". In: *Bioinformatics* 22.20, p. 2532.

Sampson, Timothy R et al. (2016). "Gut microbiota regulate motor deficits and neuroinflammation in a model of Parkinson's disease". In: *Cell* 167.6, p. 1469.

Savage, Dwayne C (1977). "Microbial ecology of the gastrointestinal tract". In: *Annual Reviews in Microbiology* 31.1, p. 107.

Sayin, Sama I et al. (2013). "Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist". In: *Cell Metabolism* 17.2, p. 225.

Scheike, Thomas H., Klaus K. Holst, and Jacob B. Hjelmborg (2013). "Estimating heritability for cause specific mortality based on twin studies". In: *Lifetime Data Analysis*, p. 1.

Schloss, Patrick D (2010). "The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies". In: *PLoS Computational Biology* 6.7, e1000844.

— (2016). "Application of a database-independent approach to assess the quality of operational taxonomic unit picking methods". In: *mSystems* 1.2, e00027–16.

Schloss, Patrick D and Jo Handelsman (2005). "Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness". In: *Applied and Environmental microbiology* 71.3, p. 1501.

Schloss, Patrick D and Sarah L Westcott (2011). "Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis". In: *Applied and Environmental Microbiology* 77.10, p. 3219.

Schloss, Patrick D et al. (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities". In: *Applied and Environmental Microbiology* 75.23.

Scholz, Matthias et al. (2016). "Strain-level microbial epidemiology and population genomics from shotgun metagenomics". In: *Nature Methods* 13.5, p. 435.

Segata, Nicola et al. (2012). "Metagenomic microbial community profiling using unique clade-specific marker genes". In: *Nature Methods* 9.8, p. 811.

Seldeen, KL, M Pang, and BR Troen (2015). "Mouse models of frailty: an emerging field". In: *Current Osteoporosis Reports* 13.5, p. 280.

Sender, Ron, Shai Fuchs, and Ron Milo (2016). "Revised estimates for the number of human and bacteria cells in the body". In: *PLoS Biology* 14.8, e1002533.

Seto, Charlie T et al. (2014). "Prolonged use of a proton pump inhibitor reduces microbial diversity: implications for Clostridium difficile susceptibility". In: *Microbiome* 2.1, p. 42.

Shankar, Vijay et al. (2015). "The networks of human gut microbe–metabolite associations are different between health and irritable bowel syndrome". In: *The ISME Journal* 9.8, p. 1899.

Shannon, Claude E (2001). "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1, p. 3.

Sharpton, Thomas J (2014). "An introduction to the analysis of shotgun metagenomic data". In: *Frontiers in Plant Science* 5, p. 209.

Shi, Na et al. (2017). "Interaction between the gut microbiome and mucosal immune system". In: *Military Medical Research* 4.1, p. 14.

Shin, Na-Ri et al. (2013). "An increase in the Akkermansia spp. population induced by metformin treatment improves glucose homeostasis in diet-induced obese mice". In: *Gut* 63, p. 706.

Smith, Patrick M et al. (2013). "The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis". In: *Science* 341.6145, p. 569.

Sneath, Peter HA and Robert R Sokal (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification.*

Soga, Tomoyoshi et al. (2003). "Quantitative metabolome analysis using capillary electrophoresis mass spectrometry". In: *Journal of Proteome Research* 2.5, p. 488.

Sogin, Mitchell L et al. (2006). "Microbial diversity in the deep sea and the underexplored "rare biosphere"". In: *Proceedings of the National Academy of Sciences* 103.32, p. 12115.

Sokol, Harry et al. (2008). "Faecalibacterium prausnitzii is an anti-inflammatory commensal bacterium identified by gut microbiota analysis of Crohn disease patients". In: *Proceedings of the National Academy of Sciences* 105.43, p. 16731.

Song, Han et al. (2016). "Faecalibacterium prausnitzii subspecies–level dysbiosis in the human gut microbiome underlying atopic dermatitis". In: *Journal of Allergy and Clinical Immunology* 137.3, p. 852.

Sonnenburg, Justin L and Fredrik Bäckhed (2016). "Diet-microbiota interactions as moderators of human metabolism". In: *Nature* 535.7610, p. 56.

Stackebrandt, EaBMG and BM Goebel (1994). "Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology". In: *International Journal of Systematic and Evolutionary Microbiology* 44.4, p. 846.

Staley, Christopher et al. (2017). "Interaction of gut microbiota with bile acid metabolism and its influence on disease states". In: *Applied Microbiology and Biotechnology* 101.1, p. 47.

Stearns, Jennifer C et al. (2011). "Bacterial biogeography of the human digestive tract". In: *Scientific Reports* 1, p. 170.

Suzuki, Ryota and Hidetoshi Shimodaira (2015). *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. R package version 2.0-0. URL: `https://CRAN.R-project.org/package=pvclust`.

Sze, Marc A and Patrick D Schloss (2016). "Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome". In: *MBio* 7.4, e01018.

Teucher, Birgit et al. (2007). "Dietary patterns and heritability of food choice in a UK female twin cohort". In: *Twin Research and Human Genetics* 10.5, p. 734.

Theriot, Casey M, Alison A Bowman, and Vincent B Young (2016). "Antibiotic-induced alterations of the gut microbiota alter secondary bile acid production and allow for Clostridium difficile spore germination and outgrowth in the large intestine". In: *MSphere* 1.1, e00045.

Thevaranjan, Netusha et al. (2017). "Age-Associated Microbial Dysbiosis Promotes Intestinal Permeability, Systemic Inflammation, and Macrophage Dysfunction". In: *Cell Host & Microbe* 21.4, p. 455.

Tomkin, Gerald H and Daphne Owens (2016). "Obesity diabetes and the role of bile acids in metabolism". In: *Journal of Translational Internal Medicine* 4.2, p. 73.

Tong, Maomeng et al. (2013). "A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease". In: *PLoS One* 8.11, e80702.

Tongeren, Sandra P van et al. (2005). "Fecal microbiota composition and frailty". In: *Applied and Environmental Microbiology* 71.10, p. 6438.

Tottey, William et al. (2013). "The human gut chip "HuGChip", an explorative phylogenetic microarray for determining gut microbiome diversity at family level". In: *PLoS One* 8.5, e62544.

Tremaroli, Valentina et al. (2015). "Roux-en-Y gastric bypass and vertical banded gastroplasty induce long-term changes on the human gut microbiome contributing to fat mass regulation". In: *Cell Metabolism* 22.2, p. 228.

Tremblay, Julien et al. (2015). "Primer and platform effects on 16S rRNA tag sequencing". In: *Frontiers in Microbiology* 6, p. 771.

Triantafyllou, Konstantinos, Christopher Chang, and Mark Pimentel (2014). "Methanogens, methane and gastrointestinal motility". In: *Journal of Neurogastroenterology and Motility* 20.1, p. 31.

Turnbaugh, Peter J et al. (2006). "An obesity-associated gut microbiome with increased capacity for energy harvest". In: *Nature* 444, p. 1027.

Turnbaugh, Peter J et al. (2008). "Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome". In: *Cell Host & Microbe* 3.4, p. 213.

Turnbaugh, Peter J et al. (2009). "A core gut microbiome in obese and lean twins". In: *Nature* 457, p. 480.

Tyson, Gene W et al. (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment". In: *Nature* 428, p. 37.

Vaishnava, Shipra et al. (2008). "Paneth cells directly sense gut commensals and maintain homeostasis at the intestinal host-microbial interface". In: *Proceedings of the National Academy of Sciences* 105.52, p. 20858.

Van Dongen, Jenny et al. (2012). "The continuing value of twin studies in the omics era". In: *Nature Reviews Genetics* 13.9, p. 640.

Väremo, Leif, Jens Nielsen, and Intawat Nookaew (2013). "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods". In: *Nucleic Acids Research* 41.8, p. 4378.

Verberkmoes, N. C. et al. (2009). "Shotgun metaproteomics of the human distal gut microbiota". In: *The ISME Journal* 3.2, p. 179.

Větrovskỳ, Tomáš and Petr Baldrian (2013). "The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses". In: *PLoS One* 8.2, e57923.

Vinolo, Marco AR et al. (2011). "Regulation of inflammation by short chain fatty acids". In: *Nutrients* 3.10, pp. 858–876.

Vos, Willem M (2013). "Fame and future of faecal transplantations–developing next-generation therapies with synthetic microbiomes". In: *Microbial Biotechnology* 6.4, p. 316.

Walker, Alan W et al. (2015). "16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice". In: *Microbiome* 3.1, p. 26.

Weingarden, Alexa R et al. (2016). "Changes in colonic bile acid composition following fecal microbiota transplantation are sufficient to control Clostridium difficile germination and growth". In: *PLoS One* 11.1, e0147210.

Weisburg, William G et al. (1991). "16S ribosomal DNA amplification for phylogenetic study." In: *Journal of Bacteriology* 173.2, p. 697.

Weiss, Sophie et al. (2016). "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision". In: *The ISME Journal* 10.7, p. 1669.

Weiss, Sophie et al. (2017). "Normalization and microbial differential abundance strategies depend upon data characteristics". In: *Microbiome* 5.1, p. 27.

Westcott, Sarah L and Patrick D Schloss (2015). "De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units". In: *PeerJ* 3, e1487.

Wildenberg, Manon E and Gijs R van den Brink (2011). "FXR activation inhibits inflammation and preserves the intestinal barrier in IBD". In: *Gut* 60, p. 432.

Williams, Ryan J, Adina Howe, and Kirsten S Hofmockel (2014). "Demonstrating microbial co-occurrence pattern analyses within and between ecosystems". In: *Frontiers in Microbiology* 5, p. 358.

Willing, Ben P et al. (2010). "A pyrosequencing study in twins shows that gastrointestinal microbial profiles vary with inflammatory bowel disease phenotypes". In: *Gastroenterology* 139.6, p. 1844.

Wolf, Yuri I, Georgy Karev, and Eugene V Koonin (2002). "Scale-free networks in biology: new insights into the fundamentals of evolution?" In: *BioEssays* 24.2, p. 105.

Wu, Gary D et al. (2011). "Linking long-term dietary patterns with gut microbial enterotypes". In: *Science* 334.6052, p. 105.

Xie, Hailiang et al. (2016). "Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome". In: *Cell Systems* 3.6, p. 572.

Yadav, Deepak, Tarini Shankar Ghosh, and Sharmila S Mande (2016). "Global investigation of composition and interaction networks in gut microbiomes of individuals belonging to diverse geographies and age-groups". In: *Gut Pathogens* 8.1, p. 17.

Yang, Bo, Yong Wang, and Pei-Yuan Qian (2016). "Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis". In: *BMC Bioinformatics* 17.1, p. 135.

Yatsunenko, Tanya et al. (2012). "Human gut microbiome viewed across age and geography". In: *Nature* 486, p. 222.

Zeevi, David et al. (2015). "Personalized nutrition by prediction of glycemic responses". In: *Cell* 163.5, p. 1079.

Zerr, Danielle M et al. (2016). "Previous antibiotic exposure increases risk of infection with extended-spectrum-β-lactamase-and AmpC-producing Escherichia coli and Klebsiella pneumoniae in pediatric patients". In: *Antimicrobial agents and chemotherapy* 60.7, p. 4237.

Zhang, Xuan et al. (2015). "The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment". In: *Nature Medicine* 21.8, p. 895.

Zhao, Ye et al. (2018). "GPR43 mediates microbiota metabolite SCFA regulation of antimicrobial peptide expression in intestinal epithelial cells via activation of mTOR and STAT3". In: *Mucosal Immunology*.

Zhernakova, Alexandra et al. (2016). "Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity". In: *Science* 352.6285, p. 565.

Ziesemer, Kirsten A et al. (2015). "Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification". In: *Scientific Reports* 5.

Appendix A

Chapter 3: Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with host frailty in the gut microbiota

The original PDF for the published manuscript accompanying Chapter 3 can be found on the accompanying digital media.

Table A.1: **S1.** Table of domains used in the construction of the frailty index.

Deficit Category	Number of separate domains
Co-morbidity	10
Physical measures	6
Biochemistry	5
Mental health	3
Self reported general health	4
Disability	8
Social functioning	1
Polypharmacy	1
Pain	1

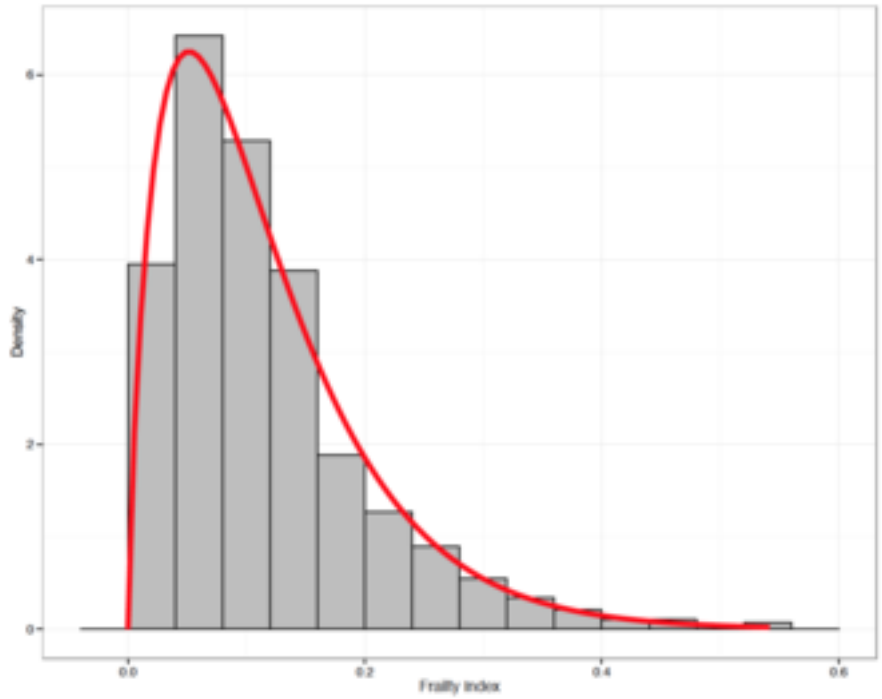


Figure A.1: **S2.** Figure showing the distribution of the FI in TwinsUK. The distribution of the frailty index in the TwinsUK cohort fits the expected gamma distribution. Histogram of untransformed FI values for the 728 individuals included in the study with a fitted gamma distribution shown in red.

Table A.2: **S3.** Table of of significant OTU-FI associations.
Due its large size, this table can be found on the accompanying digital media.

Table A.3: **S4.** Table of significant taxonomic associations with the FI.

Taxonomy	FDR Q-value Frailty vs OTU Abundance	β-Coefficient Frailty	FDR Q-value Frailty vs OTU Abundance Shannon diveristy as a covariate	β-Coefficient Frailty Shannon diveristy as a covariate
Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__[Eubacterium];s__dolichum	1.81E-04	1.99E-01	2.94E-02	1.46E-01
Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Eggerthella;s__lenta	1.81E-04	1.86E-01	5.88E-02	1.25E-01
Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Faecalibacterium;s__prausnitzii	7.99E-04	-1.82E-01	2.54E-01	-7.67E-02
Tenericutes;c__Mollicutes;o__RF39;f__;g__;s__	1.25E-03	-1.70E-01	1.87E-01	-9.47E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Lachnospira;s__	1.90E-03	-1.58E-01	2.59E-01	-7.33E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Anaerostipes;s__	6.14E-03	-1.49E-01	4.89E-01	-5.06E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__;g__;s__	6.14E-03	-1.46E-01	5.82E-01	-3.49E-02
Tenericutes;c__RF3;o__ML615J-28;f__;g__;s__	1.09E-02	-1.33E-01	3.19E-01	-7.04E-02
Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Coprobacillus;s__	2.61E-02	1.29E-01	1.87E-01	1.02E-01
Firmicutes;c__Clostridia;o__Clostridiales;f__Dehalobacteriaceae;g__Dehalobacterium;s__	3.28E-02	-1.23E-01	4.82E-01	-5.22E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__;s__	4.96E-02	-1.17E-01	7.49E-01	5.43E-03
Firmicutes;c__Clostridia;o__Clostridiales;f__Peptococcaceae;g__;s__	4.96E-02	-1.14E-01	4.97E-01	-5.04E-02
Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;g__Eggerthella	1.61E-04	1.89E-01	3.91E-02	1.29E-01
Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Faecalibacterium	8.31E-04	-1.82E-01	2.04E-01	-7.67E-02
Tenericutes;c__Mollicutes;o__RF39;f__;g__	1.16E-03	-1.70E-01	1.74E-01	-9.47E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Lachnospira	1.65E-03	-1.58E-01	2.09E-01	-7.33E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Anaerostipes	4.80E-03	-1.49E-01	3.76E-01	-5.06E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__;g__	4.80E-03	-1.46E-01	4.40E-01	-3.49E-02
Tenericutes;c__RF3;o__ML615J-28;f__;g__	8.43E-03	-1.33E-01	2.55E-01	-7.04E-02
Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Coprobacillus	1.98E-02	1.29E-01	1.74E-01	1.02E-01
Firmicutes;c__Clostridia;o__Clostridiales;f__Dehalobacteriaceae;g__Dehalobacterium	2.47E-02	-1.23E-01	3.72E-01	-5.22E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Peptococcaceae;g__	4.06E-02	-1.14E-01	3.78E-01	-5.04E-02
Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__	4.06E-02	-1.17E-01	6.20E-01	5.43E-03
Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__[Paraprevotellaceae];g__Paraprevotella	4.43E-02	-1.13E-01	2.04E-01	-8.65E-02
k__Bacteria;p__Tenericutes;c__Mollicutes;o__RF39;f__	2.12E-03	-1.70E-01	1.66E-01	-9.47E-02
k__Bacteria;p__Tenericutes;c__Mollicutes;o__Anaeroplasmatales;f__Anaeroplasmataceae	4.66E-03	-1.50E-01	1.57E-01	-1.08E-01

Table A.4: **S5.** Table of OTU associations with FI in non-parametric analyses.
Due its large size, this table can be found on the accompanying digital media.

Table A.5: **S6.** Table of OTU module assignments.
Due its large size, this table can be found on the accompanying digital media.

Appendix B

Chapter 4: Typical approaches to microbiota analyses using 16S rRNA gene sequencing: Associations with proton pump inhibitor use in the gut microbiota

The original PDF for the published manuscript accompanying Chapter 4 can be found on the accompanying digital media.

SUPPLEMENTARY METHODS

Twin pair analyses

70 pairs of MZ twins, within the 1827 individuals, were identified that were discordant for PPI use. Comparison of OTUs and collapsed taxonomies between the PPI users and non-users within pairs was carried out using two-tailed paired Wilcoxon signed rank tests on log transformed relative abundances. The resultant p-values were FDR corrected using the q-value package in R with a significance threshold of 5%. Abundances were not adjusted for any covariates in these analyses.

Site specificity of families in the Human Microbiome Project

Pre-processed 16S data was obtained from the Human Microbiome Project (HMP) in the form an OTU table.[1] This was collapsed at the family level using QIIME. Sample sites were renamed to broader categories of nose, mouth, throat, skin, vagina and gut (Supplementary Table S7). Principle components analysis was then carried out on all samples using log transformed relative abundances of families (Supplementary Figure S8). This showed at the family level there was broad overlap of throat and mouth, and skin and nose samples; therefore these groups were combined resulting in four sites for comparison.

Family abundances at each site were then compared using Kruskall-Wallis one-way ANOVA tests in R. Resultant p-values were Bonferroni adjusted and any families with $p < 0.05$ were considered to have significantly different distributions across sites. Ad hoc pair-wise comparisons were carried out on significant families using the Nemenyi test from the PMCMR package.[2] Where there were between site differences with $p < 0.05$, boxplots were used to manually assign site specificity of a family, assigning multiple sites where necessary.

Mixed effects models were made with family abundance as the response with sample site as a random effect using the coefficient of each family with each site as a measure of its site based association. To assess how the site based association of a family related to its association with PPI use, the coefficients of families at each site in the HMP data were correlated against the coefficient of families with PPI use in the TwinsUK data using Pearson correlations in R. This included 64 families that were found in both sets.

SUPPLEMENTARY LEGENDS

S1. Summary of the aspects covered by the health domains considered in the construction of the frailty index.

S2. Table of alpha diversity associations with PPI use after adjustment for covariates.

S3. Significant results from modelling OTU associations with PPI use

S4. Significant taxa associations with PPI use and replication results in the interventional data set.

S5. Significant results from mixed effects modelling of PPI and OTU and taxa associations in the subset of individuals with complete covariate data and no antibiotic use within the previous month.

S6. Table of HMP derived site specificities for collapsed families.

S7. Table showing HMP site definition mappings to the broader categories used in this study.

S8. Plot of HMP samples by PC1 and PC2 of PCA from relative abundances of all collapsed families. Samples are coloured by site. This clustering was used to group sites into the four categories: gut, mouth/throat, skin/nose and vagina.

SUPPLEMENTARY REFERENCES

1 The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* 2012;**486**:207–14. doi:10.1038/nature11234

2 Pohlert T. PMCMR: Calculate Pairwise Multiple Comparisons of Mean Rank Sums. 2015.

Table B.1: **S1.** Summary of the aspects covered by the health domains considered in the construction of the frailty index.

Deficit Category	**Number of separate domains**
Co-morbidity	10
Physical measures	6
Biochemistry	5
Mental health	3
Self reported general health	4
Disability	8
Social functioning	1
Polypharmacy	1
Pain	1

Table B.2: **S2.** Table of alpha diversity associations with PPI use after adjustment for covariates.

Alpha Diveristy Metric	**p-value (association with PPI status)**
Shannon	0.665603084
Otu Count	0.666733534
Phylogenetic Diversity	0.643844399
Chao1	0.860395912

Table B.3: **S3.** Significant results from modelling OTU associations with PPI use.

OTU ID	FDR q-value association with PPI use	Coefficient with PPI use	Assigned Taxonomy
553611	7.67E-05	0.447435365	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium; s__
4439603	8.76E-05	0.443190369	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
4424239	0.000175521	0.425407332	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
4309301	0.000204677	0.427672795	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
191355	0.000355135	-0.381535342	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
183619	0.001067258	0.377999451	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
369635	0.003746041	-0.359817024	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
4455767	0.003746041	0.355829076	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
4425214	0.004810154	0.343607548	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
519746	0.004810154	-0.3560193	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
4433947	0.005154856	0.34673914	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
New.0.CleanUp.ReferenceOTU24722	0.005154856	0.345044921	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
4411138	0.005563489	0.349259976	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Micrococcaceae; g__Rothia; s__mucilaginosa
4439360	0.005563489	0.346567007	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
3384047	0.00670208	0.338519118	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
360636	0.008046789	-0.315747457	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
529979	0.013575962	-0.310991093	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
193233	0.01479296	0.307785704	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__uniformis
3472078	0.01479296	0.314295764	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
3943182	0.01479296	0.306231401	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
New.0.ReferenceOTU286	0.016085053	-0.292848158	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
New.7.CleanUp.ReferenceOTU10733	0.016085053	0.297014214	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides
536866	0.016218003	0.309627253	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
New.0.CleanUp.ReferenceOTU17128	0.019952362	0.29964869	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
New.0.CleanUp.ReferenceOTU19985	0.020364644	0.280077785	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
193174	0.020537928	-0.306074526	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
175535	0.021437111	0.290242053	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
563572	0.021437111	-0.290894486	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
185814	0.022060366	-0.290377136	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__; s__
178205	0.023960373	-0.283826361	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Coprococcus; s__
188735	0.023960373	0.293387225	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__

Continued on next page

Table B.3 – continued from previous page

OTU ID	FDR q-value association with PPI use	Coefficient with PPI use	Assigned Taxonomy
186772	0.026882447	-0.258806021	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Faecalibacterium; s__prausnitzii
190502	0.026882447	0.289191745	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
4256470	0.026882447	0.29494603	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__ovatus
4466275	0.026882447	-0.290490515	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__; s__
561607	0.026882447	-0.291298393	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__; s__
185607	0.028342801	-0.287857024	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
187504	0.033139133	-0.255684868	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__; s__
New.0.CleanUp.ReferenceOTU25028	0.033139133	0.278508923	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__uniformis
183804	0.033627921	-0.274323234	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
New.0.CleanUp.ReferenceOTU20294	0.033627921	0.269602876	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__distasonis
4477696	0.034325967	0.273109381	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s__parainfluenzae
4401580	0.03494	0.275161879	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
New.0.ReferenceOTU224	0.042225721	0.263769891	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__distasonis
15728	0.046032842	-0.277992354	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae
New.0.CleanUp.ReferenceOTU43393	0.046151039	-0.264506464	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
191148	0.046259232	-0.270269159	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__; s__
184037	0.046270425	-0.264421885	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
4366843	0.046270425	0.270973419	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
New.0.CleanUp.ReferenceOTU29206	0.046270425	-0.255583626	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
New.0.CleanUp.ReferenceOTU6772	0.046270425	0.262324467	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__; s__
328905	0.046654261	-0.261638897	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
181155	0.047036103	0.265365707	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia; s__
New.0.CleanUp.ReferenceOTU37003	0.047036103	0.252940441	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Rikenellaceae; g__; s__

Table B.4: **S4.** Significant taxa associations with PPI use and replication results in the interventional data set.

Level	Taxonomy	FDR q-value assocation with PPI use in TwinsUK	Coefficient association in TwinsUK	Direction in Twins UK	Coefficient, Post-PPI vs Pre-PPI in Interventional Data Set	P, Pre vs Post in Interventional Data Set
Species	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia;s__mucilaginosa	2.19E-07	0.506	pos	3.8	0.28
Species	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__;s__	2.19E-07	-0.498	neg	6.9	0.35
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__anginosus	4.87E-07	0.484	pos	-0.13	0.2
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__	5.03E-07	0.464	pos	5.0	0.02
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__;s__	3.54E-06	-0.447	neg	7.1	0.09
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Other	1.35E-03	0.342	pos	NA	NA
Species	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;Other	3.01E-03	0.312	pos	0.32	0.5
Species	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Lautropia;s__	3.01E-03	0.318	pos	Not detected	Not detected
Species	k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Desulfovibrio;s__	3.01E-03	0.314	pos	-0.12	0.16
Species	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus;Other	3.01E-03	0.326	pos	Not detected	Not detected
Species	k__Bacteria;p__Cyanobacteria;c__4C0d-2;o__YS2;f__;g__;s__	3.70E-03	-0.299	neg	Not detected	Not detected
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__	4.53E-03	0.293	pos	0.29	0.27
Species	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Scardovia;s__	5.42E-03	0.299	pos	Not detected	Not detected
Species	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus;s__parainfluenzae	5.42E-03	0.287	pos	12	0.19
Species	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;Other;Other	1.45E-02	-0.277	neg	NA	NA
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella;s__parvula	1.57E-02	0.268	pos	0.23	0.23
Species	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Holdemania;s__	2.03E-02	-0.259	neg	3.2	0.43
Species	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other;Other;Other	2.61E-02	0.235	pos	0.25	0.2
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Granulicatella;s__	2.61E-02	0.251	pos	0.1	0.02
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__;s__	2.61E-02	-0.194	neg	1.6	0.18
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Mitsuokella;s__	2.61E-02	0.250	pos	Not detected	Not detected
Species	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae;g__Cardiobacterium;s__	2.61E-02	0.248	pos	Not detected	Not detected
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Oribacterium;s__	2.70E-02	0.245	pos	Not detected	Not detected
Species	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Catenibacterium;s__	2.70E-02	0.249	pos	-0.011	0.37
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella;s__dispar	3.01E-02	0.234	pos	0.04	0.13
Species	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Ruminococcus;s__	3.63E-02	-0.214	neg	3.9	0.12

Continued on next page

Table B.4 – continued from previous page

Level	Taxonomy	FDR q-value assocation with PPI use in TwinsUK	Coefficient association in TwinsUK	Direction in Twins UK	Coefficient, Post-PPI vs Pre-PPI in Interventional Data Set	P, Pre vs Post in Interventional Data Set
Species	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Bulleidia;s__p-1630-c5	3.63E-02	0.233	pos	Not detected	Not detected
Species	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__;s__	3.63E-02	0.234	pos	Not detected	Not detected
Species	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae;g__Bacteroides;s__fragilis	4.13E-02	0.230	pos	0.25	0.46
Species	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other;Other;Other	4.14E-02	0.225	pos	Not detected	Not detected
Species	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;Other;Other	4.28E-02	0.218	pos	0.11	0.09
Genus	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__	1.82E-07	-0.498	neg	0.034	0.35
Genus	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus	3.26E-07	0.469	pos	-0.13	0.24
Genus	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia	3.11E-06	0.447	pos	7.4	0.28
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__	3.11E-06	-0.447	neg	8.3	0.09
Genus	k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Desulfovibrio	2.98E-04	0.362	pos	0.32	0.9
Genus	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Lautropia	2.88E-03	0.318	pos	Not detected	Not detected
Genus	k__Bacteria;p__Cyanobacteria;c__4C0d-2;o__YS2;f__;g__	3.34E-03	-0.299	neg	Not detected	Not detected
Genus	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Scardovia	4.36E-03	0.299	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Ruminococcus	4.36E-03	-0.274	neg	0.23	0.46
Genus	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus	4.36E-03	0.287	pos	0.03	0.19
Genus	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;Other	1.14E-02	-0.277	neg	NA	NA
Genus	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Holdemania	1.65E-02	-0.259	neg	-0.013	0.43
Genus	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other;Other	1.88E-02	0.235	pos	0.25	0.2
Genus	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Granulicatella	1.88E-02	0.251	pos	0.1	0.02
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__	1.88E-02	-0.194	neg	1.55	0.18
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Mitsuokella	1.88E-02	0.256	pos	Not detected	Not detected
Genus	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae;g__Cardiobacterium	1.88E-02	0.248	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Oribacterium	1.96E-02	0.245	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Catenibacterium	1.96E-02	0.249	pos	-0.011	0.37
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Selenomonas	2.33E-02	0.240	pos	Not detected	Not detected
Genus	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Janthinobacterium	2.62E-02	0.234	pos	NA	NA
Genus	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__	2.62E-02	0.234	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other;Other	3.11E-02	0.225	pos	Not detected	Not detected
Genus	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;Other	3.18E-02	0.218	pos	0.043	0.09

Continued on next page

Table B.4 – continued from previous page

Level	Taxonomy	FDR q-value assocation with PPI use in TwinsUK	Coefficient association in TwinsUK	Direction in Twins UK	Coefficient, Post-PPI vs Pre-PPI in Interventional Data Set	P, Pre vs Post in Interventional Data Set
Genus	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae;g__Parabacteroides	3.27E-02	0.209	pos	-0.46	0.94
Genus	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus	3.54E-02	0.211	pos	-0.26	0.24
Genus	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Actinobacillus	3.54E-02	0.220	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__;g__	3.68E-02	-0.168	neg	3	0.57
Genus	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium	3.82E-02	0.209	pos	Not detected	Not detected
Genus	k__Bacteria;p__Actinobacteria;c__Coriobacteriia;o__Coriobacteriales;f__Coriobacteriaceae;Other	4.18E-02	-0.190	neg	0.05	0.56
Genus	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae;g__Weissella	4.18E-02	0.211	pos	Not detected	Not detected
Genus	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella	4.18E-02	0.201	pos	0.44	0.13
Genus	k__Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Leptotrichia	4.70E-02	0.204	pos	Not detected	Not detected
Family	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae	7.59E-07	0.458	pos	7.1	0.03
Family	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae	1.05E-06	0.464	pos	8.4	0.08
Family	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae	4.21E-04	-0.346	neg	0.22	0.08
Family	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae	7.45E-04	-0.260	neg	-0.04	0.86
Family	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae	7.45E-04	-0.331	neg	-0.03	0.84
Family	k__Bacteria;p__Cyanobacteria;c__4C0d-2;o__YS2;f__	2.15E-03	-0.299	neg	Not detected	Not detected
Family	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae	3.34E-03	0.286	pos	0.23	0.06
Family	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae	4.50E-03	0.289	pos	0.075	<0.01
Family	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other	1.76E-02	0.235	pos	0.1	0.81
Family	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae	1.76E-02	0.248	pos	Not detected	Not detected

Continued on next page

Table B.4 – continued from previous page

Level	Taxonomy	FDR q-value assocation with PPI use in TwinsUK	Coefficient association in TwinsUK	Direction in Twins UK	Coefficient, Post-PPI vs Pre-PPI in Interventional Data Set	P, Pre vs Post in Interventional Data Set
Family	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae	2.97E-02	0.231	pos	1.6	0.01
Family	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other	3.28E-02	0.225	pos	0.22	0.16
Family	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae	4.07E-02	0.209	pos	0.2	0.02
Family	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae	4.07E-02	0.207	pos	Not detected	Not detected
Family	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__	4.07E-02	-0.168	neg	0.05	0.57
Order	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales	8.52E-08	-0.400	neg	0.23	0.87
Order	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales	1.25E-06	0.434	pos	-0.037	0.22
Order	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales	6.29E-04	0.325	pos	8.8	0.09
Order	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales	6.29E-04	-0.331	neg	2.7	0.84
Order	k__Bacteria;p__Cyanobacteria;c__4C0d-2;o__YS2	2.04E-03	-0.299	neg	Not detected	Not detected
Order	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales	3.09E-03	0.286	pos	-0.031	0.19
Order	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales	1.99E-02	0.248	pos	Not detected	Not detected
Order	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales	3.71E-02	0.203	pos	0.23	0.79
Class	k__Bacteria;p__Firmicutes;c__Clostridia	3.63E-08	-0.400	neg	0.034	0.87
Class	k__Bacteria;p__Firmicutes;c__Bacilli	1.40E-06	0.415	pos	-0.037	0.22
Class	k__Bacteria;p__Firmicutes;c__Erysipelotrichi	3.39E-04	-0.331	neg	8.31	0.84
Class	k__Bacteria;p__Cyanobacteria;c__4C0d-2	1.09E-03	-0.299	neg	Not detected	Not detected
Class	k__Bacteria;p__Bacteroidetes;c__Bacteroidia	2.54E-02	0.203	pos	-0.031	0.79
Class	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria	2.63E-02	0.212	pos	0.034	0.2
Phylum	k__Bacteria;p__Firmicutes	5.82E-08	-0.383	neg	-0.07	0.71
Phylum	k__Bacteria;p__Cyanobacteria	2.67E-03	-0.271	neg	0.055	0.37

Continued on next page

Table B.4 – continued from previous page

Level	Taxonomy	FDR q-value assocation with PPI use in TwinsUK	Coefficient association in TwinsUK	Direction in Twins UK	Coefficient, Post-PPI vs Pre-PPI in Interventional Data Set	P, Pre vs Post in Interventional Data Set
Phylum	k__Bacteria;p__Bacteroidetes	1.12E-02	0.203	pos	0.031	0.79

Table B.5: **S5.** Significant results from mixed effects modelling of PPI and OTU and taxa associations in the subset of individuals with complete covariate data and no antibiotic use within the previous month.

Level	FDR q-value association with PPI use	Coefficient with PPI use	Taxonomy
OTU-4424239	0.0002	0.5640	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
OTU-553611	0.0002	0.5687	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g__Bifidobacterium; s__
OTU-519746	0.0016	-0.5189	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
OTU-New.0.CleanUp.ReferenceOTU24722	0.0018	0.5045	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-4309301	0.0023	0.5051	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
OTU-4439603	0.0023	0.4906	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
OTU-3943182	0.0079	0.4618	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-191355	0.0097	-0.4314	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
OTU-New.0.CleanUp.ReferenceOTU20294	0.0142	0.4267	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__distasonis
OTU-186352	0.0193	0.4162	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-3583645	0.0193	0.4391	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-563572	0.0193	-0.4200	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-New.0.CleanUp.ReferenceOTU17128	0.0217	0.4121	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-New.0.CleanUp.ReferenceOTU25028	0.0217	0.4088	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__uniformis
OTU-New.0.ReferenceOTU8	0.0217	0.4092	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-New.0.CleanUp.ReferenceOTU2179	0.0312	-0.3912	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__; g__; s__
OTU-184037	0.0376	-0.3857	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-3472078	0.0376	0.3970	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-4425214	0.0376	0.3723	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
OTU-4477696	0.0376	0.3856	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s__parainfluenzae
OTU-New.0.CleanUp.ReferenceOTU5423	0.0376	0.3906	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__fragilis
OTU-New.0.ReferenceOTU272	0.0413	0.3795	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-369635	0.0465	-0.3731	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
OTU-New.7.CleanUp.ReferenceOTU10733	0.0465	0.3871	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides
OTU-183619	0.0467	0.3732	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-185607	0.0467	-0.3766	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-190502	0.0467	0.3591	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__

Continued on next page

Table B.5 – continued from previous page

Level	FDR q-value association with PPI use	Coefficient with PPI use	Taxonomy
OTU-190913	0.0467	0.3584	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-191913	0.0467	-0.3727	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-193233	0.0467	0.3777	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__uniformis
OTU-4332078	0.0467	0.3316	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__; s__
OTU-4401580	0.0467	0.3765	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-4439360	0.0467	0.3655	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-4455767	0.0467	0.3511	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus; s__
OTU-New.0.CleanUp.ReferenceOTU20273	0.0467	0.3575	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__
OTU-New.0.CleanUp.ReferenceOTU34218	0.0467	-0.3788	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-New.0.CleanUp.ReferenceOTU44600	0.0467	0.3630	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__
OTU-New.0.ReferenceOTU220	0.0467	0.3616	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g__Parabacteroides; s__
OTU-184342	0.0472	-0.3597	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
OTU-New.0.ReferenceOTU286	0.0472	-0.3436	k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__; s__
OTU-175535	0.0485	0.3650	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
OTU-4433947	0.0485	0.3659	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__
SPECIES	0.0000	-0.5954	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__;s__
SPECIES	0.0000	0.5811	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia;s__mucilaginosa
SPECIES	0.0001	0.5434	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__anginosus
SPECIES	0.0002	-0.5193	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__;s__
SPECIES	0.0003	0.4684	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;s__
SPECIES	0.0045	0.4316	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Lautropia;s__
SPECIES	0.0048	0.4126	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus;s__parainfluenzae
SPECIES	0.0062	0.3999	k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Desulfovibrio;s__
SPECIES	0.0121	0.3689	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__Bulleidia;s__p-1630-c5
SPECIES	0.0179	0.3471	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus;Other
SPECIES	0.0179	0.3513	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;Other;Other
SPECIES	0.0235	0.3608	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Bifidobacterium;s__
SPECIES	0.0235	0.3368	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__
SPECIES	0.0280	0.3286	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other;Other;Other

Continued on next page

Table B.5 – continued from previous page

Level	FDR q-value association with PPI use	Coefficient with PPI use	Taxonomy
SPECIES	0.0280	0.3306	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other;Other;Other
SPECIES	0.0280	0.3347	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus;s__
SPECIES	0.0397	-0.3212	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia;s__
SPECIES	0.0397	0.3217	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella;s__parvula
SPECIES	0.0397	0.3190	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus;Other
SPECIES	0.0400	0.3077	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Oribacterium;s__
SPECIES	0.0400	0.2963	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Trabulsiella;Other
SPECIES	0.0435	-0.3086	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus;s__
SPECIES	0.0435	0.2705	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae;g__Cardiobacterium;s__
SPECIES	0.0470	0.2984	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Veillonella;s__dispar
GENUS	0.0000	-0.5954	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__
GENUS	0.0002	0.4980	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus
GENUS	0.0002	-0.5193	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae;g__
GENUS	0.0003	0.4857	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia
GENUS	0.0028	0.4316	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Lautropia
GENUS	0.0028	0.4186	k__Bacteria;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae;g__Desulfovibrio
GENUS	0.0033	0.4101	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus
GENUS	0.0082	0.4066	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae;g__Mitsuokella
GENUS	0.0135	0.3513	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;Other
GENUS	0.0234	0.3286	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other;Other
GENUS	0.0234	-0.3417	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Blautia
GENUS	0.0236	0.3306	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other;Other
GENUS	0.0354	0.3077	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Oribacterium
GENUS	0.0354	-0.2897	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae;g__Ruminococcus
GENUS	0.0354	0.2963	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Trabulsiella
GENUS	0.0390	-0.3048	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus
GENUS	0.0390	0.2705	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae;g__Cardiobacterium
FAMILY	0.0002	0.4918	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae
FAMILY	0.0002	-0.5051	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae

Continued on next page

Table B.5 – continued from previous page

Level	FDR q-value association with PPI use	Coefficient with PPI use	Taxonomy
FAMILY	0.0002	0.4891	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae
FAMILY	0.0029	0.4143	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae
FAMILY	0.0030	0.4244	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae
FAMILY	0.0063	-0.3902	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae
FAMILY	0.0211	0.3286	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other
FAMILY	0.0211	0.3306	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other
FAMILY	0.0389	0.3034	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae
FAMILY	0.0389	0.2705	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae
ORDER	0.0002	0.4712	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales
ORDER	0.0002	-0.3957	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales
ORDER	0.0006	0.4222	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales
ORDER	0.0009	0.4143	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales
ORDER	0.0023	-0.3902	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales
ORDER	0.0195	0.2705	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales
ORDER	0.0411	0.2602	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales
ORDER	0.0411	0.2787	k__Bacteria;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales
ORDER	0.0484	0.2606	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales
ORDER	0.0484	0.2518	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;Other
CLASS	0.0002	0.4712	k__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales
CLASS	0.0002	-0.3957	k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales
CLASS	0.0006	0.4222	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales
CLASS	0.0009	0.4143	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales
CLASS	0.0023	-0.3902	k__Bacteria;p__Firmicutes;c__Erysipelotrichi;o__Erysipelotrichales
CLASS	0.0195	0.2705	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales
CLASS	0.0411	0.2602	k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales
CLASS	0.0411	0.2787	k__Bacteria;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales
CLASS	0.0484	0.2606	k__Bacteria;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales
CLASS	0.0484	0.2518	k__Bacteria;p__Proteobacteria;c__Betaproteobacteria;Other
PHYLUM	0.0000	-0.3935	k__Bacteria;p__Firmicutes

Continued on next page

Table B.5 – continued from previous page

Level	FDR q-value association with PPI use	Coefficient with PPI use	Taxonomy
PHYLUM	0.0007	-0.2920	k__Bacteria;p__Cyanobacteria
PHYLUM	0.0011	0.2602	k__Bacteria;p__Bacteroidetes
PHYLUM	0.0050	-0.1445	k__Bacteria;p__Lentisphaerae
PHYLUM	0.0050	0.1579	k__Bacteria;p__Proteobacteria
PHYLUM	0.0050	-0.1643	k__Bacteria;p__Verrucomicrobia
PHYLUM	0.0050	0.1499	k__Bacteria;p__Actinobacteria
PHYLUM	0.0068	0.1096	Unassigned;Other
PHYLUM	0.0068	0.1114	k__Bacteria;p__Fusobacteria
PHYLUM	0.0070	-0.0933	k__Archaea;p__Euryarchaeota
PHYLUM	0.0070	0.0952	k__Bacteria;p__TM7
PHYLUM	0.0088	0.0655	k__Bacteria;p__Synergistetes
PHYLUM	0.0101	-0.0428	k__Bacteria;Other
PHYLUM	0.0113	0.0213	k__Bacteria;p__Tenericutes

Table B.6: **S6.** Table of HMP derived site specificities for collapsed families.

HMP Family	HMP Site Specificity (NA = None)
Root;Other;Other;Other;Other	gut
Root;p__Acidobacteria;c__;o__;f__	NA
Root;p__Acidobacteria;c__Acidobacteria;o__Acidobacteriales;f__	NA
Root;p__Acidobacteria;c__Chloracidobacteria;o__;f__	NA
Root;p__Acidobacteria;c__Holophagae;o__Holophagales;f__Holophagaceae	NA
Root;p__Acidobacteria;c__iii1-8;o__DS-18;f__	NA
Root;p__Acidobacteria;c__Solibacteres;o__Solibacterales;f__Solibacteraceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;f__CL500-29	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;f__EB1017	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;f__Iamiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;f__Microthrixaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Acidimicrobiales;Other	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__ACK-M1	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae	mouth/throat
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinosynnemataceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Bogoriellaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Brevibacteriaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Cellulomonadaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Dermabacteraceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Dermacoccaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Dietziaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Geodermatophilaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Gordoniaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Intrasporangiaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Jonesiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Kineosporiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Microbacteriaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae	mouth/throat
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micromonosporaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Mycobacteriaceae	skin/nose, vagina
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Nakamurellaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Nocardiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Nocardioidaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Promicromonosporaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Pseudonocardiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Streptomycetaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Williamsiaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Yaniellaceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;Other	skin/nose
Root;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__	gut
Root;p__Actinobacteria;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae	vagina
Root;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;f__	gut
Root;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;f__Coriobacteriaceae	mouth/throat
Root;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;Other	NA
Root;p__Actinobacteria;c__Actinobacteria;o__koll13;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__MC47;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Rubrobacterales;f__Rubrobacteraceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Solirubrobacterales;f__	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Solirubrobacterales;f__Patulibacteraceae	NA
Root;p__Actinobacteria;c__Actinobacteria;o__Solirubrobacterales;f__Solirubrobacteraceae	NA

Continued on next page

Table B.6 – continued from previous page

HMP Family	HMP Site Specificity (NA = None)
Root;p__Actinobacteria;c__Actinobacteria;o__Solirubrobacterales;Other	NA
Root;p__Armatimonadetes;c__5B-18;o__;f__	NA
Root;p__Armatimonadetes;c__Armatimonadia;o__Armatimonadales;f__Armatimonadaceae	NA
Root;p__Bacteroidetes;c__;o__;f__	NA
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__	gut
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Bacteroidaceae	gut
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae	gut, mouth/throat
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae	mouth/throat
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenellaceae	gut
Root;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;Other	gut
Root;p__Bacteroidetes;c__Flavobacteria;o__;f__	NA
Root;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Blattabacteriaceae	NA
Root;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Cryomorphaceae	NA
Root;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;f__Flavobacteriaceae	mouth/throat
Root;p__Bacteroidetes;c__Flavobacteria;o__Flavobacteriales;Other	NA
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__	skin/nose
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__Cyclobacteriaceae	NA
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__Flammeovirgaceae	NA
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__Flexibacteraceae	skin/nose
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__Saprospiraceae	NA
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;f__Sphingobacteriaceae	NA
Root;p__Bacteroidetes;c__Sphingobacteria;o__Sphingobacteriales;Other	NA
Root;p__Bacteroidetes;Other;Other;Other	mouth/throat
Root;p__BRC1;c__PRR-11;o__;f__	NA
Root;p__Chlorobi;c__OPB56;o__;f__	NA
Root;p__Chloroflexi;c__Anaerolineae;o__A31;f__S47	NA
Root;p__Chloroflexi;c__Anaerolineae;o__A4b;f__	NA
Root;p__Chloroflexi;c__Anaerolineae;o__Caldilineales;f__	NA
Root;p__Chloroflexi;c__Anaerolineae;o__Caldilineales;f__Caldilineaceae	NA
Root;p__Chloroflexi;c__Anaerolineae;Other;Other	NA
Root;p__Chloroflexi;c__Chloroflexi;Other;Other	NA
Root;p__Chloroflexi;c__SOGA31;o__;f__	NA
Root;p__Chloroflexi;c__Thermomicrobia;o__HN1-15;f__	NA
Root;p__Chloroflexi;Other;Other;Other	NA
Root;p__Cyanobacteria;c__4C0d-2;o__mle1-12;f__	NA
Root;p__Cyanobacteria;c__4C0d-2;o__YS2;f__	NA
Root;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__Chlamydomonadaceae	NA
Root;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__Trebouxiophyceae	NA
Root;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;Other	NA
Root;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__	skin/nose
Root;p__Cyanobacteria;c__Nostocophycideae;o__Nostocales;f__Nostocaceae	NA
Root;p__Cyanobacteria;c__Nostocophycideae;o__Nostocales;f__Rivulariaceae	NA
Root;p__Cyanobacteria;c__Nostocophycideae;o__Nostocales;Other	NA
Root;p__Cyanobacteria;c__Oscillatoriophycideae;o__;f__	NA
Root;p__Cyanobacteria;c__Oscillatoriophycideae;o__Chroococcales;f__Xenococcaceae	NA
Root;p__Cyanobacteria;c__Oscillatoriophycideae;Other;Other	NA
Root;p__Cyanobacteria;c__Synechococcophycideae;o__Pseudanabaenales;f__Pseudanabaenaceae	NA
Root;p__Cyanobacteria;Other;Other;Other	NA
Root;p__Elusimicrobia;c__Elusimicrobia;o__Elusimicrobiales;f__	NA
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Alicyclobacillaceae	skin/nose
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae	skin/nose
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Listeriaceae	NA
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Paenibacillaceae	NA
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Planococcaceae	NA
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae	skin/nose
Root;p__Firmicutes;c__Bacilli;o__Bacillales;f__Thermoactinomycetaceae	NA

Continued on next page

Table B.6 – continued from previous page

HMP Family	HMP Site Specificity (NA = None)
Root;p__Firmicutes;c__Bacilli;o__Bacillales;Other	skin/nose
Root;p__Firmicutes;c__Bacilli;o__Exiguobacterales;f__Exiguobacteraceae	NA
Root;p__Firmicutes;c__Bacilli;o__Gemellales;f__Gemellaceae	mouth/throat
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Aerococcaceae	mouth/throat
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae	mouth/throat
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae	NA
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae	vagina
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Leuconostocaceae	NA
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae	mouth/throat
Root;p__Firmicutes;c__Bacilli;o__Lactobacillales;Other	vagina
Root;p__Firmicutes;c__Bacilli;o__Turicibacterales;f__Turicibacteraceae	gut
Root;p__Firmicutes;c__Bacilli;Other;Other	vagina
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__	gut
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Catabacteriaceae	gut
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae	gut
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__ClostridialesFamilyXI.IncertaeSedis	mouth/throat, skin/nose, vagina
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__ClostridialesFamilyXIII.IncertaeSedis	mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Dehalobacteriaceae	gut
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Eubacteriaceae	mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae	gut, mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptococcaceae	mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptostreptococcaceae	mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Ruminococcaceae	gut
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Symbiobacteriaceae	NA
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Veillonellaceae	mouth/throat
Root;p__Firmicutes;c__Clostridia;o__Clostridiales;Other	gut
Root;p__Firmicutes;c__Clostridia;o__OPB54;f__	NA
Root;p__Firmicutes;c__Clostridia;Other;Other	NA
Root;p__Firmicutes;Other;Other;Other	gut
Root;p__Fusobacteria;c__Fusobacteria;o__Fusobacteriales;f__Fusobacteriaceae	mouth/throat
Root;p__Gemmatimonadetes;c__Gemmatimonadetes;o__Gemmatimonadales;f__	NA
Root;p__Gemmatimonadetes;c__Gemmatimonadetes;o__Gemmatimonadales;f__Gemmatimonadaceae	NA
Root;p__GN02;c__VC12-cl04;o__;f__	mouth/throat
Root;p__GN02;Other;Other;Other	NA
Root;p__Lentisphaerae;c__Lentisphaerae;o__Victivallales;f__Victivallaceae	gut
Root;p__Lentisphaerae;c__Lentisphaerae;o__Victivallales;Other	NA
Root;p__Nitrospirae;c__Nitrospira;o__Nitrospirales;f__Nitrospiraceae	NA
Root;p__Planctomycetes;c__Phycisphaerae;o__;f__	NA
Root;p__Planctomycetes;c__Phycisphaerae;o__Phycisphaerales;f__	NA
Root;p__Planctomycetes;c__Planctomycea;o__Gemmatales;f__Gemmataceae	NA
Root;p__Planctomycetes;c__Planctomycea;o__Gemmatales;f__Isosphaeraceae	NA
Root;p__Planctomycetes;c__Planctomycea;o__Pirellulales;f__	NA
Root;p__Planctomycetes;c__Planctomycea;o__Pirellulales;f__Pirellulaceae	NA
Root;p__Planctomycetes;c__Planctomycea;o__Pirellulales;Other	NA
Root;p__Planctomycetes;c__Planctomycea;o__Planctomycetales;f__Planctomycetaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__;f__	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Caulobacterales;f__Caulobacteraceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Brucellaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylobacteriaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylocystaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;Other	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Acetobacteraceae	NA

Continued on next page

Table B.6 – continued from previous page

HMP Family	HMP Site Specificity (NA = None)
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Rhodospirillaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__Rickettsiaceae	NA
Root;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae	skin/nose
Root;p__Proteobacteria;c__Alphaproteobacteria;Other;Other	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__;f__	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__	skin/nose
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Alcaligenaceae	gut
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae	mouth/throat, skin/nose
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae	skin/nose
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae	vagina
Root;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;Other	mouth/throat
Root;p__Proteobacteria;c__Betaproteobacteria;o__Gallionellales;f__Gallionellaceae	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Hydrogenophilales;f__Hydrogenophilaceae	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Methylophilales;f__Methylophilaceae	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae	mouth/throat, skin/nose
Root;p__Proteobacteria;c__Betaproteobacteria;o__Nitrosomonadales;f__Nitrosomonadaceae	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Rhodocyclales;f__	NA
Root;p__Proteobacteria;c__Betaproteobacteria;o__Rhodocyclales;f__Rhodocyclaceae	NA
Root;p__Proteobacteria;c__Betaproteobacteria;Other;Other	gut, mouth/throat
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Bdellovibrionales;f__Bdellovibrionaceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__CTD005-82B-02;f__	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfobacterales;f__Desulfobulbaceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfovibrionales;f__Desulfovibrionaceae	gut
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfuromonadales;f__Geobacteraceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Desulfuromonadales;Other	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__MIZ46;f__	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;f__	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;f__Cystobacteraceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;f__Haliangiaceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;f__Myxococcaceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;f__Polyangiaceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Myxococcales;Other	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;o__Syntrophobacterales;f__Syntrophobacteraceae	NA
Root;p__Proteobacteria;c__Deltaproteobacteria;Other;Other	NA
Root;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;f__Campylobacteraceae	mouth/throat
Root;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;f__Helicobacteraceae	NA
Root;p__Proteobacteria;c__Epsilonproteobacteria;o__Campylobacterales;Other	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Aeromonadales;f__Aeromonadaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Aeromonadales;f__Succinivibrionaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Alteromonadales;f__Shewanellaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Cardiobacteriales;f__Cardiobacteriaceae	mouth/throat
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;f__	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;f__Ectothiorhodospiraceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;f__Sinobacteraceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Chromatiales;Other	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae	skin/nose
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Legionellales;f__Coxiellaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Legionellales;f__Legionellaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Alcanivoracaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Alteromonadaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae	mouth/throat

Continued on next page

Table B.6 – continued from previous page

HMP Family	HMP Site Specificity (NA = None)
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae	mouth/throat, skin/nose
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae	skin/nose, vagina
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Thiotrichales;f__	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Vibrionales;f__Vibrionaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Sinobacteraceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae	NA
Root;p__Proteobacteria;c__Gammaproteobacteria;Other;Other	NA
Root;p__Proteobacteria;Other;Other;Other	NA
Root;p__SC3;c__;o__;f__	NA
Root;p__SC4;c__;o__;f__	NA
Root;p__SPAM;c__0319-6G9;o__;f__	NA
Root;p__Spirochaetes;c__Leptospirae;o__Leptospirales;f__Leptospiraceae	NA
Root;p__Spirochaetes;c__Spirochaetes;o__Spirochaetales;f__Spirochaetaceae	mouth/throat
Root;p__Spirochaetes;Other;Other;Other	mouth/throat
Root;p__SR1;c__;o__;f__	mouth/throat
Root;p__Synergistetes;c__Synergistia;o__Synergistales;f__Dethiosulfovibrionaceae	mouth/throat
Root;p__Synergistetes;c__Synergistia;o__Synergistales;Other	NA
Root;p__Tenericutes;c__Erysipelotrichi;o__Erysipelotrichales;f__Erysipelotrichaceae	gut, mouth/throat
Root;p__Tenericutes;c__Erysipelotrichi;o__Erysipelotrichales;f__vadinHA31	NA
Root;p__Tenericutes;c__ML615J-28;o__;f__	gut
Root;p__Tenericutes;c__Mollicutes;o__Acholeplasmatales;f__Acholeplasmataceae	NA
Root;p__Tenericutes;c__Mollicutes;o__Anaeroplasmatales;f__Anaeroplasmataceae	NA
Root;p__Tenericutes;c__Mollicutes;o__Mycoplasmatales;f__Mycoplasmataceae	vagina
Root;p__Tenericutes;c__Mollicutes;o__RF39;f__	gut
Root;p__Tenericutes;c__Mollicutes;Other;Other	NA
Root;p__Tenericutes;Other;Other;Other	NA
Root;p__Thermi;c__Deinococci;o__Deinococcales;f__Deinococcaceae	skin/nose
Root;p__TM6;c__SJA-4;o__;f__	NA
Root;p__TM7;c__TM7-1;o__;f__	NA
Root;p__TM7;c__TM7-3;o__CW040;f__	mouth/throat
Root;p__TM7;c__TM7-3;o__CW040;f__F16	mouth/throat
Root;p__TM7;c__TM7-3;o__CW040;Other	NA
Root;p__TM7;c__TM7-3;o__EW055;f__	mouth/throat, skin/nose
Root;p__TM7;c__TM7-3;o__I025;f__Rs-045	mouth/throat
Root;p__TM7;c__TM7-3;Other;Other	mouth/throat
Root;p__TM7;Other;Other;Other	NA
Root;p__Verrucomicrobia;c__Opitutae;o__Opitutales;f__Opitutaceae	NA
Root;p__Verrucomicrobia;c__Opitutae;o__Puniceicoccales;f__Puniceicoccaceae	NA
Root;p__Verrucomicrobia;c__Opitutae;Other;Other	NA
Root;p__Verrucomicrobia;c__Spartobacteria;o__Spartobacteriales;f__Spartobacteriaceae	NA
Root;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__	NA
Root;p__Verrucomicrobia;c__Verrucomicrobiae;o__Verrucomicrobiales;f__Verrucomicrobiaceae	gut
Root;p__Verrucomicrobia;Other;Other;Other	NA
Root;p__WPS-2;c__;o__;f__	NA

Table B.7: **S7.** Table showing HMP site definition mappings to the broader categories used in this study.

HMP Body Sub-Site	Collapsed Site
Stool	Gut
Attached_Keratinized_gingiva	Mouth
Buccal_mucosa	Mouth
Hard_palate	Mouth

Continued on next page

Table B.7 – continued from previous page

HMP Body Sub-Site	Collapsed Site
Saliva	Mouth
Subgingival_plaque	Mouth
Supragingival_plaque	Mouth
Tongue_dorsum	Mouth
Anterior_nares	Nose
Left_Antecubital_fossa	Skin
Left_Retroauricular_crease	Skin
Right_Antecubital_fossa	Skin
Right_Retroauricular_crease	Skin
Palatine_Tonsils	Throat
Throat	Throat
Mid_vagina	Vagina
Posterior_fornix	Vagina
Vaginal_introitus	Vagina

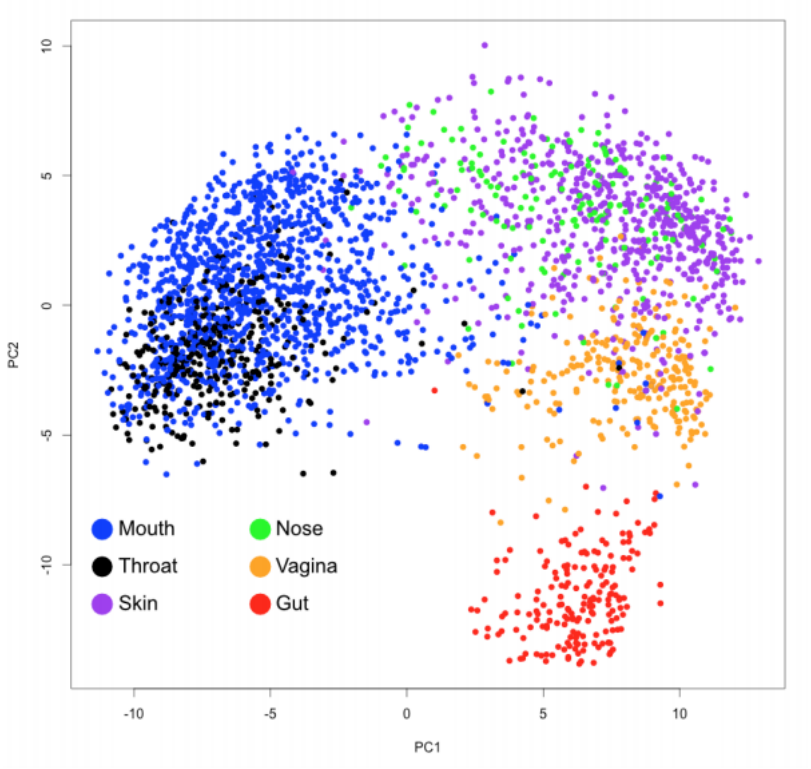


Figure B.1: **S8.** Plot of HMP samples by PC1 and PC2 of PCA from relative abundances of all collapsed families. Samples are coloured by site. This clustering was used to group sites into the four categories: gut, mouth/throat, skin/nose and vagina.

Appendix C

Chapter 5: A comparison of methods to cluster 16S rRNA gene sequences to OTUs using heritability as a measure of quality

The original PDF for the published manuscript accompanying Chapter 5 can be found on the accompanying digital media.

Overview of clustering approaches:

UCLUST De Novo (Edgar 2010)
The UCLUST algorithm developed by Robert Edgar parses a sequence list in the order of the given input (by default in QIIME this order is from most to least abundant). Starting from this first sequence UCLUST aims to generate centroids, represented by a seed or centroid sequence, from these data such that each centroids sequence is less than the % identity threshold similar to all the other centroid sequences, and that all the sequences within the centroid are greater or equal to the % identity threshold in similarity to the seed sequence. % Identity is based on global sequence alignments (considering full read lengths of all sequences in the alignment). Reads are not discarded, as any not matching to an existing centroid will become a new centroid. Identification of % distances, is carried out using a kmer based approach within a algorithm called USEARCH.

In this algorithm, all the experimental reads are broken down into kmers or 'words' of a given length. Each unique sequence in the data set is split into all possible words of the chosen length and a list of all words observed in the dataset is recorded. The number of shared kmers between reads is used as an approximate measure of sequence similarity. Once a threshold of word similarity is met, those sequences within the threshold are used in a global alignment and sequences falling within the required % identity threshold are grouped as a cluster.

UCLUST Reference
This algorithm is the same as the UCLUST de novo clustering; however instead of taking the experimental reads as input for centroids and for clustered sequences, a reference database of 16S reads is used to generate the centroids and then the experimental sequencing reads are assigned to the reference based centroids. Reads not aligning to reference centroids are discarded.

UCLUST Open Reference (Navas-Molina et al. 2013)
The open reference approach is offered as part of the QIIME wrapper. In this approach, the reads are clustered against the reference using UCLUST to generate typical closed reference OTUs. The discarded reads not matching the reference are then subjected to UCLUST de novo clustering. The resultant OTU tables are then merged to generate a complete table with closed and de novo OTUs with no reads discarded.

VSEARCH Abundance & Distance Based (Rognes et al. 2016)
VSEARCH is an open source alternative to the commercial USEARCH package (which incorporates UCLUST). The VSEARCH algorithm follows a similar kmer based heuristic approach to generate centroids as USEARCH. However, a VSEARCH implements a Needleman-Wunsch dynamic programming approach to full global alignment distances, where as by deafult USEARCH calculates heuristic distances.

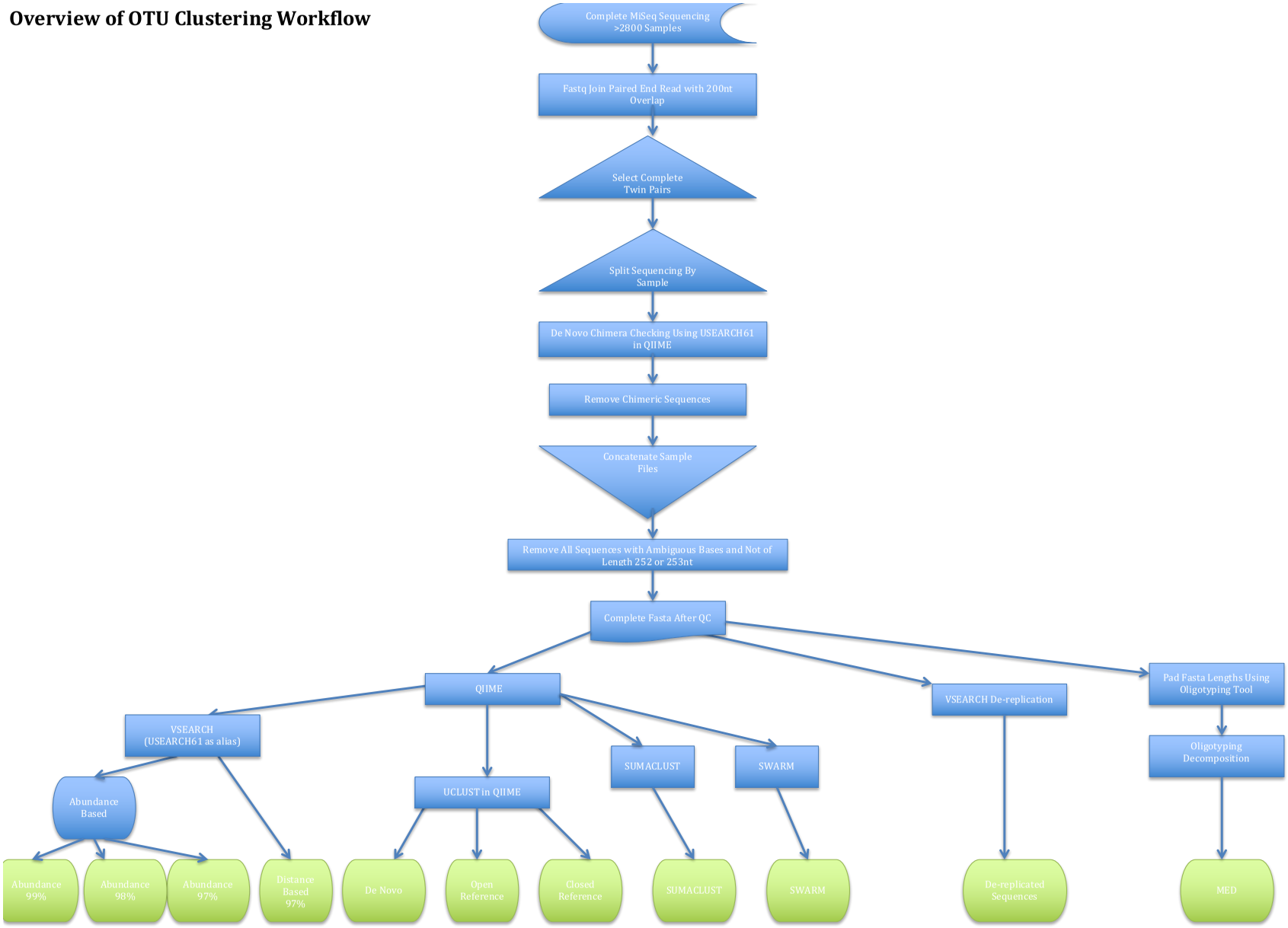
The abundance-based approach differs from the distance based where there is a tie (a sequence is > the % similar to multiple groups). In the abundance-based, it is broken by placing the sequence into the cluster that is most abundant (has the most sequences in the dataset) and in the distance-based by placing it into the cluster whose centroid shares the highest identity.

SUMACLUST (Mercier et al. 2013)
SUMACLUST is an open source algorithm that approaches clustering similarly to UCLUST. Sequences are parsed in input order (defaults are from highest to lowest abundance) and used to generate centroids matching % identity criteria. Also considering the abundance ratio between the query sequence and centroid sequence to determine if a new cluster should be made (two highly abundance sequences are more likely to be biologically distinct even is very similar in sequence). Kmer based searching is used to find similar sequences as in other heuristic clustering methods. SUMACLUST implements an alternate % identity distance that is measured as the length of the Longest Common Subsequence in an alignment divided by the shortest alignment in the region.

Swarm (Mahé et al. 2014)
Swarm is single-linkage based de novo clustering algorithm that does not utilise a single global identity threshold for clustering. When comparing sequence lengths it uses a kmer based filtering step to remove the requirement for all pair-wise distance caluclations. It then calculates Needleman-Wunsch distances between sequences determined to be similar based on kmer comparisons. The kmer threshold is set by default to only consider sequences differing by 1 nt. This identifies closely related sequences to be clustered. Swarm will start at one sequence and group outwards adding sequences to a cluster then finding sequences closely related to that one and so on. This creates chains of sequences linking together OTUs. OTUs are defined by walking the network of links and determining the abundance patterns across the network. Where drop offs in abundance are identified, chains are broken to delineate OTUs from the complete structure. In this way OTU definitions are dynamic, based on the local structure of the sequence and abundance data at each OTU in the data set.

Minimum Entropy Decomposition (A. M. Eren et al. 2014)
Minimum entropy decomposition (MED) is a de novo approach to clustering that does not directly utilise any sequence identity measures. It is based on the oligotyping approach to increasing OTU resolution (a M. Eren et al. 2014). All sequences in the data set must be aligned if reads are of variable length but can be used directly if they cover the same complete primer region. The number of A,T,C and G bases at each position across all sequences is counted. Shannon entropy, a measure of the nucleotides variability, is then calculated for each position. The assumption is that biological differences will be less randomly distributed than sequencing errors. The Shannon entropy is used as a measure of the distribution of variation at each position. The sequences are then split into smaller groups based on their nucleotide at the highest entropy position. This is carried out recursively on each split group, splitting into further groups by their base at each most variable position until the total entropy values no longer meet predetermined thresholds. If a group falls below a minimum abundance threshold its reads are discarded.

Overview of OTU Clustering Workflow



Commands to remove chimeric sequences & trim data (run on individual FASTA files for each sample):

Identify chimeras (QIIME):
identify_chimeric_seqs.py -i sample.fasta -m usearch61 --suppress_usearch61_ref --usearch61_xn 7 -o sample.out

Concatenate the resultant chimera files into one file then remove chimeras from full data FASTA using (QIIME):
filter_fasta.py -f full.fasta -o no_chimeras.fasta -s concat_chimeras.txt –n

Trim to length and remove any with ambiguous bases (Mothur):
trim.seqs(fasta=no_chimeras.fasta, maxambig=0, minlength=252, maxlength=253)

The resultant file was then used as the input.fasta in the steps below.

Commands to cluster OTUs for each method:

QIIME commands were run using QIIME version 1.9.0, where VSEARCH was required USEARCH61 was run as a method within QIIME and an alias used to direct USEARCH61 calls to the VSEARCH version 1.9.3 (Linux) executable.

VSEARCH de-replication was carried out directly using VSEARCH following a workflow provided by Greg Caporaso at https://gist.github.com/gregcaporaso/f3c042e5eb806349fa18 (last accessed 20th April 2016).

Minimum Entropy Decomposition (MED) was run using the decompose command as part of the oliogtyping pipeline, details of which can be found here: http://merenlab.org/projects/oligotyping/

Picking parameters either matched those of the He et al. comparison (He et al. 2015) where applicable or used default values. Where a reference was required Greengenes version 13_8 (DeSantis et al. 2006).

UCLUST DE NOVO:

pick_de_novo_otus.py -i input.fasta -o uclust_denovo/ -p uclustdenovo.params.txt

uclustdenovo.params.txt:
pick_otus:otu_picking_method uclust
pick_otus:max_accepts 16
pick_otus:max_rejects 64
pick_otus:enable_rev_strand_match True

UCLUST CLOSED REFERENCE:

pick_closed_reference_otus.py -i input.fasta -o uclust_closed_gg_13_8/ -r gg_13_8_otus/rep_set/97_otus.fasta –p closedref.params.uclust.txt

closedref.params.uclust.txt:
pick_otus:otu_picking_method uclust_ref
pick_otus:max_accepts 16
pick_otus:max_rejects 64
pick_otus:enable_rev_strand_match True
pick_otus:minlen 30

UCLUST OPEN REFERENCE:

pick_open_reference_otus.py -i input.fasta -o uclust_open_gg_13_8/ -m uclust -r gg_13_8_otus/rep_set/97_otus.fasta -s 1 -p openref.params.uclust.txt --min_otu_size 1 --prefilter_percent_id 0.0

openref.params.uclust.txt:
pick_otus:max_accepts 16
pick_otus:max_rejects 64
pick_otus:enable_rev_strand_match True

VSEARCH AGC 97, 98 and 99%:
pick_de_novo_otus.py -i input.fasta -o vsearch_AGC_**% threshold**/ -p agc.params.**% Threshold**.txt

agc.params files:
pick_otus:otu_picking_method usearch61
pick_otus:sizeorder True
pick_otus:max_accepts 16
pick_otus:max_rejects 64
pick_otus:enable_rev_strand_match True
pick_otus:minlen 30
pick_otus:similarity 0.97**0.98**0.99

VSEARCH DGC 97%:
pick_de_novo_otus.py -i input.fasta -o vsearch_DGC_97/ -p dgc.params.txt

dgc.params.txt:
pick_otus:otu_picking_method usearch61
pick_otus:max_accepts 16
pick_otus:max_rejects 64
pick_otus:enable_rev_strand_match True
pick_otus:minlen 30

SUMACLUST:
pick_de_novo_otus.py -i input.fasta -o sumaclust_denovo/ -p $ sumaclust.params.txt

sumaclust.params.txt:
pick_otus:otu_picking_method sumaclust
pick_otus:threads 16

SWARM:
pick_de_novo_otus.py -i input.fasta -o swarm_denovo/ -p swarm.params.txt

swarm.params.txt:
pick_otus:otu_picking_method swarm
pick_otus:threads 16

VSEARCH DE-REPLICATION:
vsearch --derep_input.fasta --output vsearch_dereplicated/vsearch_dereplicated_repset.fna --uc vsearch_dereplicated/vsearch_dereplicated.uc --relabel_sha1 --relabel_keep

biom from-uc -i vsearch_dereplicated/vsearch_dereplicated.uc -o vsearch_dereplicated/vsearch_dereplicated_biom.biom --rep-set-fp vsearch_dereplicated/vsearch_dereplicated_repset.fna

MED:
o-pad-with-gaps input.fasta -o input_padded.fasta
decompose –i input_padded.fasta –o med/

Heritability Calculation

Total OTU counts were taken from complete OTU tables (biom summarize-table) then tables subset to OTUs found in at least 50% of individuals. This enabled the smaller tables to be converted into text formats to load in R whilst retaining complete count information, which was used to convert counts to relative abundances.

R Heritability Script is based on the OpenMX path based ACE modelling example (http://openmx.psyc.virginia.edu/docs/OpenMx/2.5.1/GeneticEpi_Path.html accessed last on 30/6/2016).

Heritability R script:

```
#packages for compostional transforms, lme4 and for open mx
library(OpenMx)
library(lme4)
library(data.table)
library(methods)
library(car)

#read in argument from command line
args = commandArgs(trailingOnly=TRUE)
#otu table for otus in >= 50% samples
otutable=args[1]
#output directory
out=args[2]
#path to OTU counts for samples in complete (all OTUs not just >50%) OTU table
summarypath=args[3]
#total number of OTUs observed in the compelete set
totobs=as.numeric(args[4])

#read in OTU table
bacteria=fread(otutable,header=T,sep="\t")
bacteria=data.frame(bacteria)
rownames(bacteria)=bacteria[,1]
```

```
bacteria=bacteria[,-1]

#transform data to abundances considering total number of OTUs and adding a pseduo count of
1 for all OTUs
bacteria1=bacteria
summary=read.table(summarypath)
summary=summary[match(colnames(bacteria1),summary[,1]),]
for(i in 1:nrow(summary)){
  summary[i,2]=summary[i,2]+totobs
}
#summary now represents the total number of OTUs in each sample in the complete OTU table,
plus a count of 1 on every OTU
#add a count of 1 to the zeros on the OTU table
bacteria1=bacteria1+1
#convert to relative abundances
for(i in 1:ncol(bacteria1)){
  bacteria1[,i]=bacteria1[,i]/summary[i,2]
}
#clrb used from now on to hold otu counts
clrb=bacteria1


#read in mapping file
sampdat=read.csv("twins_mapping_file.csv",header=T)
#read in file summarising the read depth in each sample
sampseqcount=read.table("twins_seq_count.txt",header=T,sep='\t')
sampdat=merge(sampdat,sampseqcount,by="X.SampleID")

#create box cox transformed residuals after controlling for covariates

#subset the data to the twins in the study from the mapping and OTU tables make all the
files the same order
clrb=t(clrb)
clrb=clrb[which(rownames(clrb)%in%sampdat$X.SampleID),]
sampdat=sampdat[which(sampdat$X.SampleID%in%rownames(clrb)),]
sampdat=sampdat[match(rownames(clrb),sampdat$X.SampleID),]
#clrbresids will hold the residuals
clrbresids=clrb

#scale covariates from mapping to same ranges
#age
sage=scale(sampdat$age_at_sample,center = T)
#sequencing depth
sscount=scale(sampdat$SeqCount,center=T)
gender=as.factor(sampdat$i.gender)
#sequencing run
seqrun=as.factor(sampdat$p.SequencingRun)
#who loaded and extracted the data
loaded=as.factor(sampdat$e.Loadedby)
extracted=as.factor(sampdat$e.Extractedby)
#sample collected in person or from postal kit
collect=as.factor(sampdat$s.CollectionMethod)

#generate box cox transformed residuals otu by otu
for(i in 1:ncol(clrb)){
  otu=clrb[,i]
  mod=lm(otu~gender+sage+seqrun+sscount+loaded+extracted+collect)
  #estimate box cox lamda and normalise
  a=powerTransform(otu)
  mod=lm(bcPower(otu,a$roundlam)~gender+sage+seqrun+sscount+loaded+extracted+collect)
  resids=summary(mod)$residuals
  #overwrite the column with residuals
  clrbresids[,i]=resids
}


#create twin format for the openMX ACE script
fam=c()
twin1=c()
twin2=c()
zyg=c()
for(i in unique(sampdat$i.FamilyID)){
  dat=sampdat[which(sampdat$i.FamilyID==i),]
  fam=c(fam,i)
  twin1=c(twin1,which(sampdat$X.SampleID==dat$X.SampleID[1]))
  twin2=c(twin2,which(sampdat$X.SampleID==dat$X.SampleID[2]))
  zyg=c(zyg,as.character(dat$i.zygosity[1]))
```

```
}

t1dz=twin1[which(zyg=="DZ")]
t2dz=twin2[which(zyg=="DZ")]
t1mz=twin1[which(zyg=="MZ")]
t2mz=twin2[which(zyg=="MZ")]
fdz=fam[which(zyg=="DZ")]
fmz=fam[which(zyg=="MZ")]

#vectors to hold estimates and CI
a=c()
al=c()
au=c()

c=c()
cl=c()
cu=c()


e=c()
el=c()
eu=c()

#calculate hertiabilties for each OTU reformatting data
for(i in 1:ncol(clrbresids)){
  otudat=clrbresids[,i]
  otunamer=colnames(clrbresids)[i]

  #get data for pairs
  otu1dz=otudat[t1dz]
  otu2dz=otudat[t2dz]
  otu1mz=otudat[t1mz]
  otu2mz=otudat[t2mz]

  #group
  mzData=cbind(otu1mz,otu2mz)
  dzData=cbind(otu1dz,otu2dz)

  #colnames to sel vars
  selVars=c("otut1","otut2")
  colnames(mzData)=selVars
  colnames(dzData)=selVars

  #define latent ace variables
  aceVars=c("A1","C1","E1","A2","C2","E2")

  #means and covs
  mzmeans=colMeans(mzData,na.rm=TRUE)
  dzmeans=colMeans(dzData,na.rm=TRUE)
  mzcov=cov(mzData,use="complete")
  dzcov=cov(dzData,use="complete")

  # Path objects for Multiple Groups
  manifestVars=selVars
  latentVars=aceVars
  # variances of latent variables
  latVariances <- mxPath( from=aceVars, arrows=2,
                          free=FALSE, values=1 )
  # means of latent variables
  latMeans     <- mxPath( from="one", to=aceVars, arrows=1,
                          free=FALSE, values=0 )
  # means of observed variables
  obsMeans     <- mxPath( from="one", to=selVars, arrows=1,
                          free=TRUE, values=0, labels="mean" )
  # path coefficients for twin 1
  pathAceT1    <- mxPath( from=c("A1","C1","E1"), to="otut1", arrows=1,
                          free=TRUE, values=.33,  label=c("a","c","e") )
  # path coefficients for twin 2
  pathAceT2    <- mxPath( from=c("A2","C2","E2"), to="otut2", arrows=1,
                          free=TRUE, values=.33,  label=c("a","c","e") )
  # covariance between C1 & C2
  covC1C2      <- mxPath( from="C1", to="C2", arrows=2,
                          free=FALSE, values=1 )
  # covariance between A1 & A2 in MZ twins
  covA1A2_MZ   <- mxPath( from="A1", to="A2", arrows=2,
                          free=FALSE, values=1 )
  # covariance between A1 & A2 in DZ twins
```

```
  covA1A2_DZ   <- mxPath( from="A1", to="A2", arrows=2,
                          free=FALSE, values=.5 )

  # Data objects for Multiple Groups
  dataMZ       <- mxData( observed=mzData, type="raw" )
  dataDZ       <- mxData( observed=dzData, type="raw" )

  # Combine Groups
  paths        <- list( latVariances, latMeans, obsMeans,
                        pathAceT1, pathAceT2, covC1C2 )
  modelMZ      <- mxModel(model="MZ", type="RAM", manifestVars=selVars,
                          latentVars=aceVars, paths, covA1A2_MZ, dataMZ )
  modelDZ      <- mxModel(model="DZ", type="RAM", manifestVars=selVars,
                          latentVars=aceVars, paths, covA1A2_DZ, dataDZ )
  minus2ll     <- mxAlgebra( expression=MZ.fitfunction + DZ.fitfunction,
                             name="minus2loglikelihood" )
  obj          <- mxFitFunctionAlgebra( "minus2loglikelihood" )

  estVA <- mxAlgebra(expression=a*a,name="a2")
  estVC <- mxAlgebra(expression=c*c,name="c2")
  estVE <- mxAlgebra(expression=e*e,name="e2")
  estVP <- mxAlgebra(expression=a2+c2+e2, name="V")
  estPropVA <- mxAlgebra(expression=a2/V, name="A")
  estPropVC <- mxAlgebra(expression=c2/V, name="C")
  estPropVE <- mxAlgebra(expression=e2/V, name="E")


  modelACE  <- mxModel(model="ACE", modelMZ, modelDZ, minus2ll,obj,
estVA,estVC,estVE,estVP,estPropVA,estPropVC,estPropVE,mxCI(c("A","C","E")))

  # Run Model
  fitACE       <- mxRun(modelACE,intervals=TRUE)

  a2l=fitACE$output$confidenceIntervals[1,1]
  a2=fitACE$output$confidenceIntervals[1,2]
  a2u=fitACE$output$confidenceIntervals[1,3]

  c2l=fitACE$output$confidenceIntervals[2,1]
  c2=fitACE$output$confidenceIntervals[2,2]
  c2u=fitACE$output$confidenceIntervals[2,3]

  e2l=fitACE$output$confidenceIntervals[3,1]
  e2=fitACE$output$confidenceIntervals[3,2]
  e2u=fitACE$output$confidenceIntervals[3,3]

  a=c(a,a2)
  al=c(al,a2l)
  au=c(au,a2u)

  c=c(c,c2)
  cl=c(cl,c2l)
  cu=c(cu,c2u)

  e=c(e,e2)
  el=c(el,e2l)
  eu=c(eu,e2u)
}

#combine results into a data frame
res=data.frame(colnames(clrbresids),al,a,au,cl,c,cu,el,e,eu)
```

References:

DeSantis, T.Z. et al., 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Applied and Environmental Microbiology*, 72(7), pp.5069–5072. Available at: http://aem.asm.org/cgi/doi/10.1128/AEM.03006-05.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19), pp.2460–2461. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq461.

Eren, a M. et al., 2014. Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, 111(28), pp.E2875–E2884. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24965363.

Eren, A.M. et al., 2014. Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, 9(4), pp.968–979. Available at: http://www.nature.com/doifinder/10.1038/ismej.2014.195.

He, Y. et al., 2015. Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity.

Microbiome, 3(1), p.20. Available at: http://www.microbiomejournal.com/content/3/1/20.
Mahé, F. et al., 2014. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2, p.e593. Available at: http://dx.doi.org/10.7717/peerj.593.
Mercier, C. et al., 2013. SUMATRA and SUMACLUST : fast and exact comparison and clustering of sequences. *Abstract In: SeqBio 25-26th Nov 2013*, p.27. Available at: http://doi.wiley.com/10.1002/ejoc.201200111.
Navas-Molina, J.A. et al., 2013. Advancing Our Understanding of the Human Microbiome Using QIIME. In *Microbial Metagenomics, Metatranscriptomics, and Metaproteomics*. Elsevier Inc., pp. 371–444. Available at: http://linkinghub.elsevier.com/retrieve/pii/B9780124078635000198.
Rognes, T. et al., 2016. VSEARCH. Available at: https://github.com/torognes/vsearch.

Table C.1: **S1.** Summary of OTU counts resulting from each method

Method	**Number of OTUs Observed**	**Table Density**	**Mean Counts/Sample**	**No. OTUs in >50% Samples**	**No. Heritable OTUs (A >mean(A) & lower 95% CI > 0.01)**
SUMACLUST (97VSEARCH Distance (97VSEARCH Abundance 97VSEARCH Abundance 98VSEARCH Abundance 99VSEARCH Dereplicated (100SWARM	3078044	0.001	81040	361	54
Minimum Entropy Decomposition	364	0.559	67247	203	43
UCLUST De Novo	864784	0.002	81040	329	66
UCLUST Closed Reference	11382	0.096	79330	667	126
UCLUST Open Reference	692047	0.003	81040	672	101

Table C.2: **S2.** Method-wise A,C and E estimates for each taxa found across all methods.

Due its large size, this table can be found on the accompanying digital media.

Appendix D

Chapter 6: Identification of taxonomic markers that define a health-associated gut microbiome

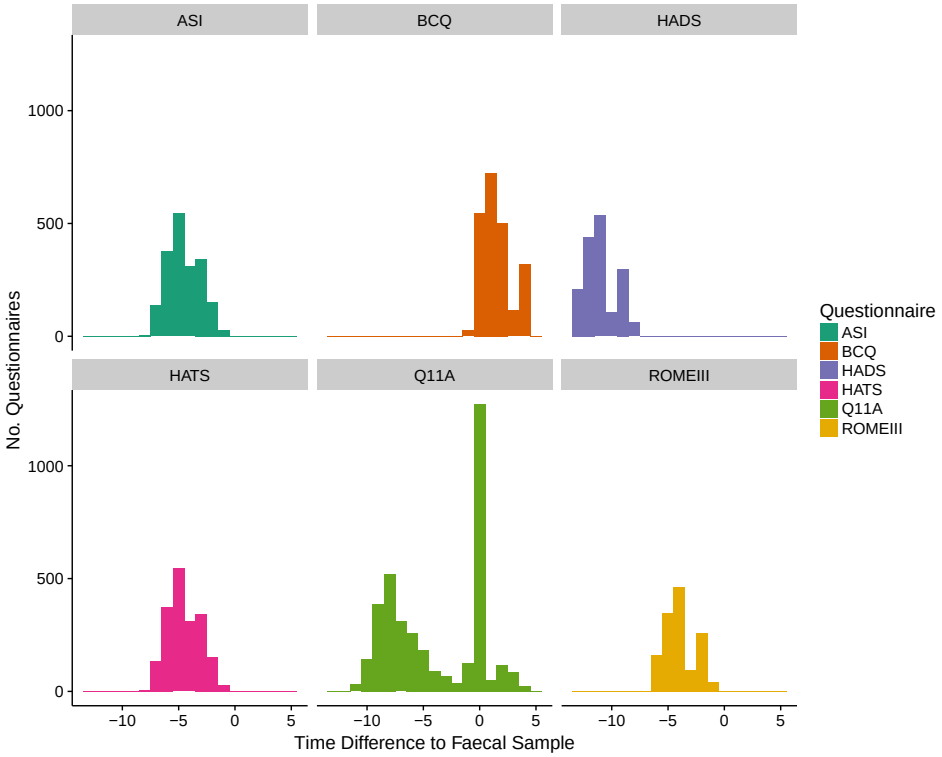


Figure D.1: Questionnaire time differences from the faecal sample dates.

Table D.1: Mapping of the 206 gut microbiome features considered to their marker in the 68 marker features used in analyses.

Gut Microbiome Trait	Representative marker trait in analysis
Weighted_UniFrac_PCoA_MDS6	Weighted_UniFrac_PCoA_MDS6
Weighted_UniFrac_PCoA_MDS5	Weighted_UniFrac_PCoA_MDS5
Weighted_UniFrac_PCoA_MDS4	Weighted_UniFrac_PCoA_MDS4
Weighted_UniFrac_PCoA_MDS3	Weighted_UniFrac_PCoA_MDS3
Weighted_UniFrac_PCoA_MDS2	Weighted_UniFrac_PCoA_MDS2
Weighted_UniFrac_PCoA_MDS1	Weighted_UniFrac_PCoA_MDS1
Unweighted_UniFrac_PCoA_MDS6	Unweighted_UniFrac_PCoA_MDS6
Unweighted_UniFrac_PCoA_MDS5	Unweighted_UniFrac_PCoA_MDS5
Unweighted_UniFrac_PCoA_MDS4	Unweighted_UniFrac_PCoA_MDS4
Unweighted_UniFrac_PCoA_MDS3	Unweighted_UniFrac_PCoA_MDS3
Unweighted_UniFrac_PCoA_MDS2	Unweighted_UniFrac_PCoA_MDS2
Unweighted_UniFrac_PCoA_MDS1	Unweighted_UniFrac_PCoA_MDS1
shannon	shannon
PD_whole_tree	PD_whole_tree
observed_otus	PD_whole_tree
k__Bacteria.p__Verrucomicrobia.c__Verrucomicrobiae.o__Verrucomicrobiales.f__Verrucomicrobiaceae	k__Bacteria.p__Verrucomicrobia.c__Verrucomicrobiae
k__Bacteria.p__Verrucomicrobia.c__Verrucomicrobiae	k__Bacteria.p__Verrucomicrobia.c__Verrucomicrobiae
k__Bacteria.p__Verrucomicrobia.c__Verruco.5.o__WCHB1.41.f__RFP12	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__Verruco.5	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__Opitutae.o__Opitutales.f__Opitutaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__Opitutae.o__.Cerasicoccales..f__.Cerasicoccaceae.	k__Bacteria.p__Verrucomicrobia.c__Opitutae.o__.Cerasicoccales..f__.Cerasicoccaceae.
k__Bacteria.p__Verrucomicrobia.c__Opitutae	k__Bacteria.p__Verrucomicrobia.c__Opitutae.o__.Cerasicoccales..f__.Cerasicoccaceae.
k__Bacteria.p__Verrucomicrobia.c__.Spartobacteria..o__.Chthoniobacterales..f__.Chthoniobacteraceae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__.Spartobacteria.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__.Pedosphaerae..o__.Pedosphaerales..f__auto67_4W	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Verrucomicrobia.c__.Pedosphaerae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Tenericutes.c__RF3	k__Bacteria.p__Tenericutes.c__RF3
k__Bacteria.p__Tenericutes.c__Mollicutes.o__Mycoplasmatales.f__Mycoplasmataceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Tenericutes.c__Mollicutes.o__Anaeroplasmatales.f__Anaeroplasmataceae	k__Bacteria.p__Tenericutes.c__Mollicutes.o__Anaeroplasmatales.f__Anaeroplasmataceae
k__Bacteria.p__Tenericutes.c__Mollicutes	k__Bacteria.p__Tenericutes.c__Mollicutes
k__Bacteria.p__Synergistetes.c__Synergistia.o__Synergistales.f__Synergistaceae	k__Bacteria.p__Synergistetes.c__Synergistia.o__Synergistales.f__Synergistaceae
k__Bacteria.p__Synergistetes.c__Synergistia.o__Synergistales.f__Dethiosulfovibrionaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Synergistetes.c__Synergistia	k__Bacteria.p__Synergistetes.c__Synergistia.o__Synergistales.f__Synergistaceae
k__Bacteria.p__Spirochaetes.c__Spirochaetes.o__Spirochaetales.f__Spirochaetaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Spirochaetes.c__Spirochaetes.o__Sphaerochaetales.f__Sphaerochaetaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Spirochaetes.c__Spirochaetes	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Spirochaetes.c__.Brachyspirae..o__.Brachyspirales..f__Brachyspiraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Spirochaetes.c__.Brachyspirae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Xanthomonadales.f__Xanthomonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Xanthomonadales.f__Sinobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Vibrionales.f__Vibrionaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Vibrionales.f__Pseudoalteromonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pseudomonadales.f__Pseudomonadaceae	k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pseudomonadales.f__Pseudomonadaceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pseudomonadales.f__Moraxellaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae	k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Pasteurellales.f__Pasteurellaceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Oceanospirillales.f__Oceanospirillaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae	k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Chromatiales.f__Chromatiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Cardiobacteriales.f__Cardiobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Alteromonadales.f__Shewanellaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Alteromonadales.f__.Chromatiaceae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Aeromonadales.f__Succinivibrionaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Aeromonadales.f__Aeromonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria	k__Bacteria.p__Proteobacteria.c__Gammaproteobacteria.o__Enterobacteriales.f__Enterobacteriaceae
k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria.o__Campylobacterales.f__Helicobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria.o__Campylobacterales.f__Campylobacteraceae	k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria
k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria	k__Bacteria.p__Proteobacteria.c__Epsilonproteobacteria
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Syntrophobacterales.f__Syntrophobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfovibrionales.f__Desulfovibrionaceae	k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfobacterales.f__Desulfobulbaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria.o__Desulfobacterales.f__Desulfobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria	k__Bacteria.p__Proteobacteria.c__Deltaproteobacteria
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Rhodocyclales.f__Rhodocyclaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Nitrosomonadales.f__Nitrosomonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Neisseriales.f__Neisseriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Methylophilales.f__Methylophilaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Oxalobacteraceae	k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Oxalobacteraceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Comamonadaceae	k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Comamonadaceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Burkholderiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Alcaligenaceae	k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Alcaligenaceae
k__Bacteria.p__Proteobacteria.c__Betaproteobacteria	k__Bacteria.p__Proteobacteria.c__Betaproteobacteria.o__Burkholderiales.f__Alcaligenaceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Sphingomonadales.f__Sphingomonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rickettsiales.f__mitochondria	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodospirillales.f__Rhodospirillaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Hyphomonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Rhizobiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Phyllobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Methylobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Hyphomicrobiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Brucellaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Bradyrhizobiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhizobiales.f__Beijerinckiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Caulobacterales.f__Caulobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria
k__Bacteria.p__Planctomycetes.c__Planctomycetia.o__Planctomycetales.f__Planctomycetaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Planctomycetes.c__Planctomycetia.o__Pirellulales.f__Pirellulaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Planctomycetes.c__Planctomycetia.o__Gemmatales.f__Gemmataceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Planctomycetes.c__Planctomycetia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Planctomycetes.c__Phycisphaerae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__OD1.c__ZB2	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Nitrospirae.c__Nitrospira.o__Nitrospirales.f__0319.6A21	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Nitrospirae.c__Nitrospira	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Lentisphaerae.c__.Lentisphaeria..o__Victivallales.f__Victivallaceae	k__Bacteria.p__Lentisphaerae.c__.Lentisphaeria.
k__Bacteria.p__Lentisphaerae.c__.Lentisphaeria.	k__Bacteria.p__Lentisphaerae.c__.Lentisphaeria.
k__Bacteria.p__Gemmatimonadetes.c__Gemm.1	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Leptotrichiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Fusobacteriaceae	k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Fusobacteriaceae

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Fusobacteria.c__Fusobacteriia	k__Bacteria.p__Fusobacteria.c__Fusobacteriia.o__Fusobacteriales.f__Fusobacteriaceae
k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae	k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae
k__Bacteria.p__Firmicutes.c__Erysipelotrichi	k__Bacteria.p__Firmicutes.c__Erysipelotrichi.o__Erysipelotrichales.f__Erysipelotrichaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Thermoanaerobacterales.f__Thermoanaerobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Natranaerobiales.f__Anaerobrancaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Veillonellaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Veillonellaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Syntrophomonadaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Syntrophomonadaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Ruminococcaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Ruminococcaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Peptostreptococcaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Peptostreptococcaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Peptococcaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Peptococcaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Lachnospiraceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Gracilibacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Eubacteriaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Eubacteriaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__EtOH8	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__EtOH8
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Dehalobacteriaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Dehalobacteriaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Clostridiaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Clostridiaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Christensenellaceae	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Christensenellaceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__Caldicoprobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__.Tissierellaceae.	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__.Tissierellaceae.
k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__.Mogibacteriaceae.	k__Bacteria.p__Firmicutes.c__Clostridia.o__Clostridiales.f__.Mogibacteriaceae.
k__Bacteria.p__Firmicutes.c__Clostridia	k__Bacteria.p__Firmicutes.c__Clostridia
k__Bacteria.p__Firmicutes.c__Bacilli.o__Turicibacterales.f__Turicibacteraceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Turicibacterales.f__Turicibacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Leuconostocaceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Leuconostocaceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Lactobacillaceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Lactobacillaceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Enterococcaceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Enterococcaceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Carnobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Aerococcaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Gemellales.f__Gemellaceae	k__Bacteria.p__Firmicutes.c__Bacilli.o__Gemellales.f__Gemellaceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Staphylococcaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Planococcaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Paenibacillaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Listeriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Bacillaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli.o__Bacillales.f__Alicyclobacillaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Firmicutes.c__Bacilli	k__Bacteria.p__Firmicutes.c__Bacilli.o__Lactobacillales.f__Streptococcaceae
k__Bacteria.p__Fibrobacteres.c__Fibrobacteria.o__Fibrobacterales.f__Fibrobacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Fibrobacteres.c__Fibrobacteria	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Elusimicrobia.c__Elusimicrobia.o__Elusimicrobiales.f__Elusimicrobiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Elusimicrobia.c__Elusimicrobia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Deferribacteres.c__Deferribacteres.o__Deferribacterales.f__Deferribacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Deferribacteres.c__Deferribacteres	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Cyanobacteria.c__Chloroplast	k__Bacteria.p__Cyanobacteria.c__Chloroplast
k__Bacteria.p__Cyanobacteria.c__4C0d.2	k__Bacteria.p__Cyanobacteria.c__4C0d.2
k__Bacteria.p__Chloroflexi.c__Ellin6529	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chloroflexi.c__Anaerolineae.o__Anaerolineales.f__Anaerolinaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chloroflexi.c__Anaerolineae.o__A31.f__S47	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chloroflexi.c__Anaerolineae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chlorobi.c__SJA.28	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chlamydiae.c__Chlamydiia.o__Chlamydiales.f__Parachlamydiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Chlamydiae.c__Chlamydiia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Sphingobacteriia.o__Sphingobacteriales.f__Sphingobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Sphingobacteriia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Flavobacteriia.o__Flavobacteriales.f__Flavobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Flavobacteriia.o__Flavobacteriales.f__.Weeksellaceae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Flavobacteriia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Cytophagia.o__Cytophagales.f__Cytophagaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Cytophagia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__S24.7	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__S24.7
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Rikenellaceae	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Rikenellaceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Prevotellaceae	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Prevotellaceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Porphyromonadaceae	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Porphyromonadaceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__p.2534.18B5	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__BS11	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Bacteroidaceae	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__Bacteroidaceae
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Paraprevotellaceae.	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Paraprevotellaceae.

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Odoribacteraceae.	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Odoribacteraceae.
k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Barnesiellaceae.	k__Bacteria.p__Bacteroidetes.c__Bacteroidia.o__Bacteroidales.f__.Barnesiellaceae.
k__Bacteria.p__Bacteroidetes.c__Bacteroidia	Weighted_UniFrac_PCoA_MDS2
k__Bacteria.p__Bacteroidetes.c__.Saprospirae..o__.Saprospirales..f__Saprospiraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__.Saprospirae..o__.Saprospirales..f__Chitinophagaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Bacteroidetes.c__.Saprospirae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Thermoleophilia.o__Solirubrobacterales.f__Conexibacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Thermoleophilia.o__Gaiellales.f__Gaiellaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Thermoleophilia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Coriobacteriia.o__Coriobacteriales.f__Coriobacteriaceae	k__Bacteria.p__Actinobacteria.c__Coriobacteriia.o__Coriobacteriales.f__Coriobacteriaceae
k__Bacteria.p__Actinobacteria.c__Coriobacteriia	k__Bacteria.p__Actinobacteria.c__Coriobacteriia.o__Coriobacteriales.f__Coriobacteriaceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Bifidobacteriales.f__Bifidobacteriaceae	k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Bifidobacteriales.f__Bifidobacteriaceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Streptomycetaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Sanguibacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Promicromonosporaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Nocardioidaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Nocardiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Mycobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Micromonosporaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Micrococcaceae	k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Micrococcaceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Microbacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Intrasporangiaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Gordoniaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Dermabacteraceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Corynebacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Cellulomonadaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Brevibacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Actinosynnemataceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Actinomycetaceae	k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Actinomycetales.f__Actinomycetaceae
k__Bacteria.p__Actinobacteria.c__Actinobacteria	k__Bacteria.p__Actinobacteria.c__Actinobacteria.o__Bifidobacteriales.f__Bifidobacteriaceae
k__Bacteria.p__Acidobacteria.c__PAUC37f	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__Acidobacteriia.o__Acidobacteriales.f__Acidobacteriaceae	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__Acidobacteriia	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae

Continued on next page

Table D.1 – continued from previous page

Gut Microbiome Trait	Representative marker trait in analysis
k__Bacteria.p__Acidobacteria.c__Acidobacteria.6.o__iii1.15.f__RB40	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__Acidobacteria.6.o__iii1.15.f__mb2424	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__Acidobacteria.6	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__.Chloracidobacteria..o__RB41.f__Ellin6075	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Bacteria.p__Acidobacteria.c__.Chloracidobacteria.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Archaea.p__Euryarchaeota.c__Thermoplasmata.o__E2.f__.Methanomassiliicoccaceae.	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Archaea.p__Euryarchaeota.c__Thermoplasmata	k__Bacteria.p__Proteobacteria.c__Alphaproteobacteria.o__Rhodobacterales.f__Rhodobacteraceae
k__Archaea.p__Euryarchaeota.c__Methanobacteria.o__Methanobacteriales.f__Methanobacteriaceae	k__Archaea.p__Euryarchaeota.c__Methanobacteria
k__Archaea.p__Euryarchaeota.c__Methanobacteria	k__Archaea.p__Euryarchaeota.c__Methanobacteria

Table D.2: Results from logistic regression analysis of disorders versus microbiome features.
Due its large size, this table can be found on the accompanying digital media.

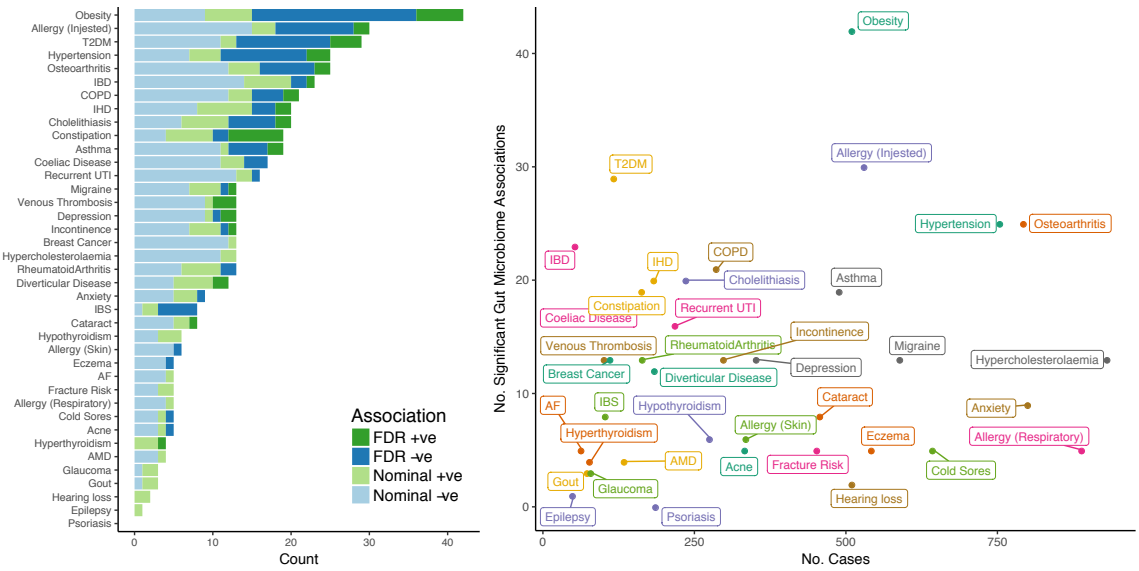


Figure D.2: Disorder associations with microbiota markers without adjustment for BMI and additionally considering obesity.

Table D.3: Results from logistic regression analysis of disorders versus microbiome features without adjustment for BMI.
Due its large size, this table can be found on the accompanying digital media.

Table D.4: Edge table for the inferred Bayesian network underlying the disorder, drug and microbiome data.
Due its large size, this table can be found on the accompanying digital media.

Table D.5: Enrichment analysis results for KEGG annotated pathways from metagenomic data against the HMI, number of disorders, and frailty.
Due its large size, this table can be found on the accompanying digital media.

Table D.6: Enrichment analysis results for KEGG annotated pathways from faecal metabolite data against the HMI, number of disorders, and frailty.

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
HMI	Lipid	52	0.91197	0.1736	0.66284	0.8264	1	18	34
HMI	Amino Acid	56	-2.9237	0.9993	0.9999	7.00E-04	0.0147	9	47
HMI	Unknown	147	-0.12484	0.5547	0.86192	0.4453	1	44	103
HMI	Fatty Acid, Dicarboxylate	10	5.0394	0	0	1	1	9	1
HMI	Lysine Metabolism	10	-1.5888	0.9484	0.9999	0.0516	0.3612	1	9
HMI	Sphingolipid Metabolism	6	-2.6439	0.997	0.9999	0.003	0.042	0	6
HMI	Endocannabinoid	4	-3.2076	0.9999	0.9999	1.00E-04	0.0042	0	4
HMI	Nucleotide	8	-0.47401	0.685	0.87182	0.315	1	1	7
HMI	Aminosugar Metabolism	7	0.77636	0.2186	0.66523	0.7814	1	3	4
HMI	Glycine, Serine And Threonine Metabolism	3	-1.0701	0.8571	0.9999	0.1429	0.75022	0	3
HMI	Xenobiotics	13	-0.2378	0.5797	0.86192	0.4203	1	3	10
HMI	Alanine And Aspartate Metabolism	7	-2.0217	0.9831	0.9999	0.0169	0.17745	0	7
HMI	Methionine, Cysteine, Sam And Taurine Metabolism	10	-1.7578	0.9654	0.9999	0.0346	0.29064	0	10
HMI	Cofactors And Vitamins	10	-0.122	0.5489	0.86192	0.4511	1	3	7
HMI	Food Component/plant	21	0.92192	0.1715	0.66284	0.8285	1	6	15
HMI	Chemical	10	1.2185	0.1085	0.50633	0.8915	1	3	7
HMI	Secondary Bile Acid Metabolism	5	0.76451	0.2184	0.66523	0.7816	1	3	2
HMI	Pentose Metabolism	7	-0.27116	0.6094	0.86192	0.3906	1	2	5
HMI	Gamma-glutamyl Amino Acid	6	0.72489	0.2389	0.66523	0.7611	1	3	3
HMI	Dipeptide	15	-0.34351	0.625	0.86192	0.375	1	2	13
HMI	Fatty Acid, Monohydroxy	4	2.2767	0.0132	0.1848	0.9868	1	3	1
HMI	Pyrimidine Metabolism, Uracil Containing	5	0.63042	0.2602	0.66523	0.7398	1	2	3
HMI	Nicotinate And Nicotinamide Metabolism	4	-1.085	0.8638	0.9999	0.1362	0.75022	0	4
HMI	Phenylalanine And Tyrosine Metabolism	13	0.24353	0.3958	0.80487	0.6042	1	4	9

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
HMI	Leucine, Isoleucine And Valine Metabolism	10	1.6705	0.049	0.343	0.951	1	5	5
HMI	Glutamate Metabolism	9	0.17573	0.4216	0.80487	0.5784	1	3	6
HMI	Histidine Metabolism	7	-0.38616	0.6567	0.86192	0.3433	1	1	6
HMI	Carbohydrate	5	0.19396	0.4212	0.80487	0.5788	1	1	4
HMI	Urea Cycle; Arginine And Proline Metabolism	5	0.31411	0.3733	0.80487	0.6267	1	2	3
HMI	Tocopherol Metabolism	3	2.7283	0.0051	0.1071	0.9949	1	3	0
HMI	Tryptophan Metabolism	6	-0.19115	0.5744	0.86192	0.4256	1	1	5
HMI	Purine Metabolism, Guanine Containing	4	-0.63606	0.7316	0.90374	0.2684	1	0	4
HMI	Phospholipid Metabolism	8	0.5539	0.2847	0.66523	0.7153	1	3	5
HMI	Tca Cycle	5	-0.36493	0.6389	0.86192	0.3611	1	1	4
HMI	Fructose, Mannose And Galactose Metabolism	4	-0.13319	0.5461	0.86192	0.4539	1	1	3
HMI	Purine Metabolism, Adenine Containing	5	2.0742	0.0217	0.22785	0.9783	1	4	1
HMI	Sterol	4	1.8023	0.0367	0.30828	0.9633	1	3	1
HMI	Lysolipid	3	0.07335	0.4556	0.83197	0.5444	1	0	3
HMI	Long Chain Fatty Acid	3	0.66065	0.2527	0.66523	0.7473	1	2	1
HMI	Pyrimidine Metabolism, Cytidine Containing	3	0.55706	0.2851	0.66523	0.7149	1	1	2
HMI	Peptide	3	1.3478	0.092	0.483	0.908	1	2	1
HMI	Pyrimidine Metabolism, Orotate Containing	3	1.5215	0.0663	0.3978	0.9337	1	3	0
Frailty	Fatty Acid, Dicarboxylate	10	-4.4944	1	1	0	0	0	10
Frailty	Lipid	52	1.5948	0.0449	0.27195	0.9551	0.9993	28	24
Frailty	Unknown	147	-1.9476	0.989	1	0.011	0.231	60	87
Frailty	Lysine Metabolism	10	1.9115	0.0253	0.21252	0.9747	0.9993	7	3
Frailty	Phenylalanine And Tyrosine Metabolism	13	1.6456	0.0481	0.27195	0.9519	0.9993	10	3

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
Frailty	Cofactors And Vitamins	10	-0.7196	0.7663	1	0.2337	0.61346	5	5
Frailty	Fatty Acid, Monohydroxy	4	-1.1307	0.8698	1	0.1302	0.49713	2	2
Frailty	Purine Metabolism, Adenine Containing	5	-1.4411	0.9248	1	0.0752	0.3948	1	4
Frailty	Food Component/plant	21	0.11964	0.4514	1	0.5486	0.92165	11	10
Frailty	Nucleotide	8	-1.4619	0.93	1	0.07	0.3948	2	6
Frailty	Aminosugar Metabolism	7	-1.6412	0.9502	1	0.0498	0.3948	0	7
Frailty	Sphingolipid Metabolism	6	3.2718	7.00E-04	0.0294	0.9993	0.9993	6	0
Frailty	Gamma-glutamyl Amino Acid	6	-1.8479	0.9682	1	0.0318	0.3948	0	6
Frailty	Secondary Bile Acid Metabolism	5	1.4851	0.0687	0.3206	0.9313	0.9993	4	1
Frailty	Chemical	10	0.073166	0.4656	1	0.5344	0.92165	5	5
Frailty	Glutamate Metabolism	9	-0.62156	0.7278	1	0.2722	0.63513	3	6
Frailty	Endocannabinoid	4	2.1274	0.014	0.147	0.986	0.9993	4	0
Frailty	Phospholipid Metabolism	8	2.2472	0.0095	0.133	0.9905	0.9993	7	1
Frailty	Pyrimidine Metabolism, Uracil Containing	5	-0.88119	0.8123	1	0.1877	0.60642	1	4
Frailty	Amino Acid	56	2.3803	0.005	0.105	0.995	0.9993	37	19
Frailty	Tca Cycle	5	-0.97045	0.8353	1	0.1647	0.57645	2	3
Frailty	Purine Metabolism, Guanine Containing	4	-0.44594	0.6705	1	0.3295	0.69195	1	3
Frailty	Dipeptide	15	-0.77517	0.7827	1	0.2173	0.61346	7	8
Frailty	Tocopherol Metabolism	3	-0.78053	0.78	1	0.22	0.61346	1	2
Frailty	Fructose, Mannose And Galactose Metabolism	4	-1.5665	0.9411	1	0.0589	0.3948	0	4
Frailty	Carbohydrate	5	0.73346	0.2314	0.6942	0.7686	0.9993	2	3
Frailty	Urea Cycle; Arginine And Proline Metabolism	5	-0.64029	0.7414	1	0.2586	0.63513	1	4
Frailty	Sterol	4	0.91671	0.179	0.6265	0.821	0.9993	3	1
Frailty	Methionine, Cysteine, Sam And Taurine Metabolism	10	1.1827	0.1188	0.4536	0.8812	0.9993	6	4
Frailty	Xenobiotics	13	1.6093	0.0518	0.27195	0.9482	0.9993	9	4

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
Frailty	Peptide	3	-1.3288	0.9086	1	0.0914	0.42653	0	3
Frailty	Long Chain Fatty Acid	3	-1.558	0.9408	1	0.0592	0.3948	0	3
Frailty	Tryptophan Metabolism	6	-0.30115	0.6251	1	0.3749	0.7498	3	3
Frailty	Alanine And Aspartate Metabolism	7	1.2316	0.1111	0.4536	0.8889	0.9993	6	1
Frailty	Pyrimidine Metabolism, Cytidine Containing	3	-1.182	0.8813	1	0.1187	0.49713	0	3
Frailty	Leucine, Isoleucine And Valine Metabolism	10	-0.2683	0.6069	1	0.3931	0.75046	3	7
Frailty	Histidine Metabolism	7	0.74937	0.2259	0.6942	0.7741	0.9993	5	2
Frailty	Pentose Metabolism	7	-0.21508	0.5877	1	0.4123	0.7529	3	4
Frailty	Nicotinate And Nicotinamide Metabolism	4	0.44864	0.3245	0.85181	0.6755	0.9993	2	2
Frailty	Pyrimidine Metabolism, Orotate Containing	3	0.29123	0.3863	0.95439	0.6137	0.99136	1	2
Frailty	Lysolipid	3	-0.51659	0.6931	1	0.3069	0.67841	1	2
Frailty	Glycine, Serine And Threonine Metabolism	3	0.46455	0.3216	0.85181	0.6784	0.9993	2	1
No. Disorders	Aminosugar Metabolism	7	-3.0926	0.9993	0.9999	7.00E-04	0.0147	2	5
No. Disorders	Fatty Acid, Dicarboxylate	10	-3.9668	0.9999	0.9999	1.00E-04	0.0042	1	9
No. Disorders	Lipid	52	2.6129	0.0029	0.1218	0.9971	0.9971	32	20
No. Disorders	Food Component/plant	21	0.41581	0.3344	0.92978	0.6656	0.98721	11	10
No. Disorders	Unknown	147	-0.8856	0.849	0.9999	0.151	0.57655	74	73
No. Disorders	Lysine Metabolism	10	1.0565	0.1462	0.53445	0.8538	0.98721	7	3

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
No. Disorders	Cofactors And Vitamins	10	-0.73259	0.7719	0.9999	0.2281	0.7287	5	5
No. Disorders	Purine Metabolism, Adenine Containing	5	-2.279	0.9883	0.9999	0.0117	0.12285	1	4
No. Disorders	Amino Acid	56	1.9378	0.0221	0.4641	0.9779	0.9971	39	17
No. Disorders	Nicotinate And Nicotinamide Metabolism	4	0.90223	0.1906	0.61578	0.8094	0.98721	3	1
No. Disorders	Phospholipid Metabolism	8	1.3026	0.0922	0.53445	0.9078	0.98721	7	1
No. Disorders	Phenylalanine And Tyrosine Metabolism	13	0.032736	0.4796	0.9999	0.5204	0.84792	9	4
No. Disorders	Chemical	10	-0.045135	0.5197	0.9999	0.4803	0.84792	6	4
No. Disorders	Xenobiotics	13	1.5391	0.0606	0.53445	0.9394	0.98721	9	4
No. Disorders	Leucine, Isoleucine And Valine Metabolism	10	-1.3972	0.9234	0.9999	0.0766	0.40215	2	8
No. Disorders	Carbohydrate	5	-0.69235	0.7571	0.9999	0.2429	0.7287	3	2
No. Disorders	Sterol	4	1.0846	0.1443	0.53445	0.8557	0.98721	2	2
No. Disorders	Tca Cycle	5	-0.75348	0.7772	0.9999	0.2228	0.7287	3	2
No. Disorders	Alanine And Aspartate Metabolism	7	1.556	0.0598	0.53445	0.9402	0.98721	5	2
No. Disorders	Endocannabinoid	4	1.1683	0.1269	0.53445	0.8731	0.98721	4	0
No. Disorders	Peptide	3	-2.4487	0.9926	0.9999	0.0074	0.1036	0	3

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
No. Disorders	Pyrimidine Metabolism, Cytidine Containing	3	-1.8248	0.9627	0.9999	0.0373	0.31332	0	3
No. Disorders	Gamma-glutamyl Amino Acid	6	-1.0483	0.855	0.9999	0.145	0.57655	3	3
No. Disorders	Glutamate Metabolism	9	-0.43018	0.6572	0.9999	0.3428	0.80033	4	5
No. Disorders	Urea Cycle; Arginine And Proline Metabolism	5	0.05942	0.4751	0.9999	0.5249	0.84792	2	3
No. Disorders	Tocopherol Metabolism	3	1.1791	0.1215	0.53445	0.8785	0.98721	3	0
No. Disorders	Tryptophan Metabolism	6	0.55022	0.2976	0.8928	0.7024	0.98721	3	3
No. Disorders	Secondary Bile Acid Metabolism	5	-0.41233	0.657	0.9999	0.343	0.80033	3	2
No. Disorders	Pentose Metabolism	7	-0.011829	0.504	0.9999	0.496	0.84792	4	3
No. Disorders	Dipeptide	15	-0.33318	0.6346	0.9999	0.3654	0.80773	8	7
No. Disorders	Fructose, Mannose And Galactose Metabolism	4	-1.1467	0.8777	0.9999	0.1223	0.57073	1	3
No. Disorders	Lysolipid	3	-0.59523	0.7214	0.9999	0.2786	0.73132	2	1
No. Disorders	Histidine Metabolism	7	1.1984	0.1181	0.53445	0.8819	0.98721	5	2
No. Disorders	Methionine, Cysteine, Sam And Taurine Metabolism	10	0.37072	0.3542	0.92978	0.6458	0.98721	6	4
No. Disorders	Nucleotide	8	-1.6675	0.952	0.9999	0.048	0.336	2	6
No. Disorders	Fatty Acid, Monohydroxy	4	-0.1643	0.5725	0.9999	0.4275	0.84792	2	2

Continued on next page

Table D.6 – continued from previous page

Phenotype	Metabolite Group	Metabolites (tot)	Stat (dist.dir)	p (postive enrichment)	p adj (postive enrichment)	p (negative enrichment)	p adj (negative enrichment)	Metabolites (positive)	Metabolites (negative)
No. Disorders	Sphingolipid Metabolism	6	1.07	0.148	0.53445	0.852	0.98721	5	1
No. Disorders	Glycine, Serine And Threonine Metabolism	3	1.0308	0.1527	0.53445	0.8473	0.98721	3	0
No. Disorders	Pyrimidine Metabolism, Uracil Containing	5	-1.4346	0.9241	0.9999	0.0759	0.40215	0	5
No. Disorders	Pyrimidine Metabolism, Orotate Containing	3	-0.28967	0.6111	0.9999	0.3889	0.81669	1	2
No. Disorders	Purine Metabolism, Guanine Containing	4	-0.057596	0.5331	0.9999	0.4669	0.84792	2	2
No. Disorders	Long Chain Fatty Acid	3	-0.63434	0.7323	0.9999	0.2677	0.73132	0	3
≥									

Appendix E

Chapter 7: Identifying stable communities in the gut microbiota with consistent associations to host phenotypes

Table E.1: OTU community assignments and heritability analysis results for the communities in TwinsUK. Due its large size, this table can be found on the accompanying digital media.

Table E.2: Community association results with age and BMI in TwinsUK (Community members can be found in Table E.1)

Community	Age Beta	Age p-value	Age FDR Adj. p-value	BMI Beta	BMI p-value	BMI FDR Adj. p-value
Module_1	0.001	0.963	0.973	0.021	0.264	0.379
Module_10	-0.040	0.036	0.069	0.015	0.433	0.533
Module_11	0.062	0.001	0.005	-0.075	0.000	0.000
Module_12	0.105	0.000	0.000	-0.126	0.000	0.000
Module_13	-0.013	0.514	0.617	0.006	0.741	0.831
Module_14	0.021	0.276	0.378	-0.108	0.000	0.000
Module_15	0.037	0.055	0.096	-0.044	0.022	0.044
Module_16	0.086	0.000	0.000	-0.050	0.009	0.021
Module_17	0.117	0.000	0.000	0.046	0.016	0.035
Module_18	-0.016	0.395	0.486	-0.020	0.310	0.413
Module_19	0.058	0.003	0.009	-0.030	0.119	0.190
Module_2	-0.050	0.009	0.024	0.007	0.718	0.821
Module_20	0.073	0.000	0.001	-0.086	0.000	0.000
Module_21	-0.038	0.048	0.085	-0.022	0.246	0.358
Module_22	0.046	0.016	0.038	-0.041	0.033	0.065
Module_23	-0.050	0.009	0.024	-0.098	0.000	0.000
Module_24	-0.120	0.000	0.000	0.034	0.072	0.126
Module_25	-0.010	0.597	0.682	-0.017	0.374	0.466
Module_26	0.071	0.000	0.001	-0.139	0.000	0.000
Module_27	-0.059	0.002	0.008	-0.071	0.000	0.001
Module_28	0.021	0.266	0.376	-0.061	0.002	0.005
Module_29	-0.046	0.017	0.039	-0.018	0.357	0.450
Module_3	-0.064	0.001	0.003	0.100	0.000	0.000
Module_30	-0.024	0.206	0.300	0.005	0.795	0.877
Module_31	-0.003	0.890	0.926	0.051	0.008	0.020

Continued on next page

Table E.2 – continued from previous page

Community	Age Beta	Age p-value	Age FDR Adj. p-value	BMI Beta	BMI p-value	BMI FDR Adj. p-value
Module_32	0.069	0.000	0.002	-0.050	0.009	0.021
Module_33	0.046	0.018	0.039	-0.050	0.009	0.021
Module_34	-0.007	0.701	0.765	-0.025	0.190	0.289
Module_35	0.058	0.002	0.008	-0.094	0.000	0.000
Module_36	0.036	0.061	0.104	-0.061	0.001	0.004
Module_37	-0.001	0.961	0.973	0.009	0.628	0.735
Module_38	0.076	0.000	0.000	-0.081	0.000	0.000
Module_39	0.069	0.000	0.002	0.035	0.070	0.125
Module_4	-0.092	0.000	0.000	-0.002	0.924	0.954
Module_40	0.027	0.158	0.237	-0.115	0.000	0.000
Module_41	0.040	0.040	0.073	0.001	0.978	0.978
Module_42	0.058	0.002	0.008	-0.124	0.000	0.000
Module_43	0.060	0.002	0.007	-0.112	0.000	0.000
Module_44	-0.017	0.365	0.461	0.050	0.008	0.020
Module_45	0.006	0.764	0.824	-0.087	0.000	0.000
Module_46	0.017	0.375	0.468	-0.039	0.044	0.085
Module_47	0.061	0.001	0.006	-0.061	0.001	0.004
Module_48	0.109	0.000	0.000	-0.077	0.000	0.000
Module_49	-0.044	0.023	0.047	-0.010	0.593	0.711
Module_5	-0.065	0.001	0.003	0.113	0.000	0.000
Module_50	-0.005	0.787	0.839	-0.020	0.296	0.406
Module_51	-0.040	0.037	0.069	0.004	0.835	0.911
Module_52	0.078	0.000	0.000	0.108	0.000	0.000
Module_53	0.032	0.097	0.150	0.031	0.109	0.178
Module_54	-0.139	0.000	0.000	0.077	0.000	0.000
Module_55	0.081	0.000	0.000	-0.129	0.000	0.000
Module_56	-0.055	0.004	0.012	0.035	0.069	0.125
Module_57	-0.061	0.001	0.006	-0.010	0.614	0.728
Module_58	-0.041	0.032	0.062	0.001	0.942	0.962
Module_59	0.010	0.594	0.682	-0.063	0.001	0.003
Module_6	0.058	0.003	0.008	0.024	0.210	0.310
Module_60	0.035	0.070	0.117	-0.058	0.002	0.007
Module_61	0.043	0.025	0.049	-0.031	0.101	0.167
Module_62	0.008	0.688	0.760	-0.003	0.885	0.938
Module_63	0.109	0.000	0.000	-0.109	0.000	0.000
Module_64	-0.028	0.143	0.218	-0.047	0.014	0.031
Module_65	0.048	0.012	0.030	-0.033	0.090	0.154
Module_66	0.014	0.478	0.581	-0.006	0.744	0.831
Module_67	0.008	0.671	0.749	-0.054	0.005	0.013
Module_68	0.111	0.000	0.000	-0.089	0.000	0.000
Module_69	-0.025	0.191	0.281	0.030	0.122	0.191
Module_7	-0.019	0.310	0.408	-0.060	0.002	0.005
Module_70	-0.033	0.080	0.131	-0.056	0.003	0.008
Module_71	0.034	0.080	0.131	-0.020	0.294	0.406
Module_72	0.032	0.092	0.145	-0.019	0.332	0.431
Module_73	0.045	0.018	0.039	-0.083	0.000	0.000
Module_74	0.046	0.018	0.039	0.002	0.897	0.938
Module_75	-0.002	0.897	0.926	0.008	0.685	0.793
Module_76	-0.134	0.000	0.000	0.002	0.899	0.938
Module_77	-0.049	0.011	0.027	0.013	0.484	0.588
Module_78	-0.011	0.553	0.648	0.032	0.092	0.155
Module_79	0.032	0.092	0.145	-0.025	0.194	0.291
Module_8	0.039	0.041	0.075	-0.080	0.000	0.000
Module_80	-0.020	0.293	0.396	-0.043	0.025	0.051
Module_81	0.050	0.009	0.024	-0.063	0.001	0.003
Module_82	-0.020	0.302	0.403	0.036	0.062	0.115
Module_83	0.056	0.004	0.010	-0.098	0.000	0.000
Module_84	0.005	0.803	0.847	-0.021	0.285	0.402

Continued on next page

Table E.2 – continued from previous page

Community	Age Beta	Age p-value	Age FDR Adj. p-value	BMI Beta	BMI p-value	BMI FDR Adj. p-value
Module_85	0.012	0.522	0.619	-0.045	0.018	0.037
Module_86	-0.010	0.607	0.685	0.038	0.046	0.086
Module_87	0.021	0.273	0.378	-0.029	0.137	0.212
Module_88	0.103	0.000	0.000	-0.019	0.324	0.425
Module_89	-0.056	0.003	0.010	-0.001	0.953	0.963
Module_9	0.045	0.018	0.039	-0.092	0.000	0.000
Module_90	0.019	0.333	0.426	0.020	0.302	0.408
Module_91	0.066	0.001	0.003	-0.067	0.000	0.001
Module_92	0.044	0.023	0.047	-0.118	0.000	0.000
Module_93	-0.019	0.328	0.426	-0.003	0.872	0.938
Module_94	0.023	0.228	0.327	-0.168	0.000	0.000
Module_95	-0.057	0.003	0.010	0.050	0.009	0.021
Module_96	0.000	0.979	0.979	-0.018	0.343	0.439

Table E.3: Table summarising the OTUs within each of the community types identified across all three datasets.
Due its layout, this table can be found on the accompanying digital media.

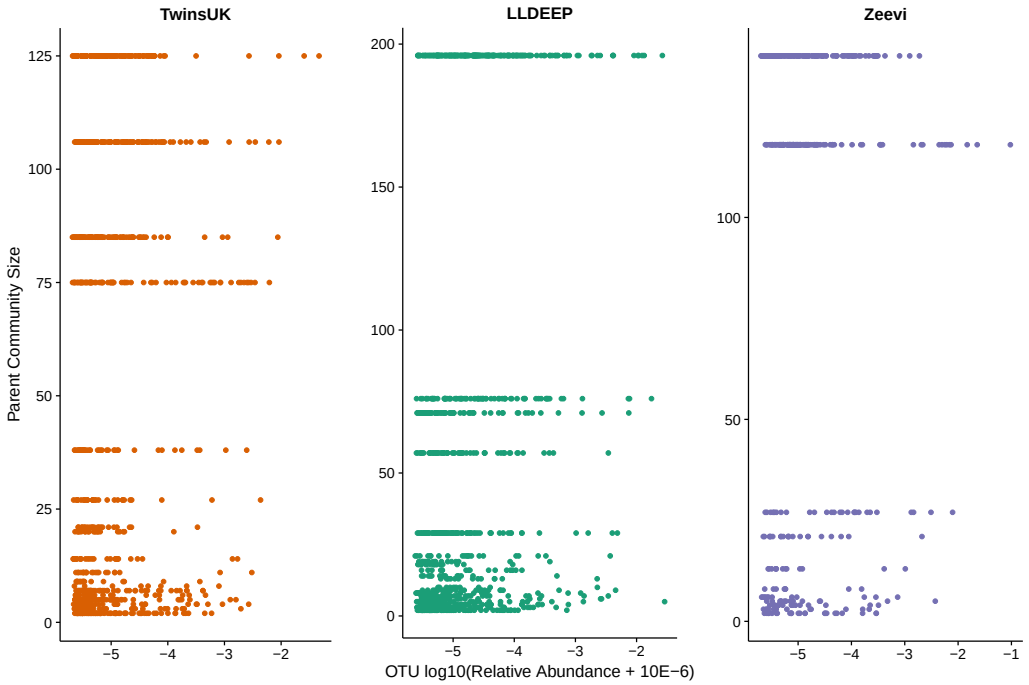


Figure E.1: A comparison of the relative abundance of an OTU to the number of OTUs in its assigned community for each of the three data sets.

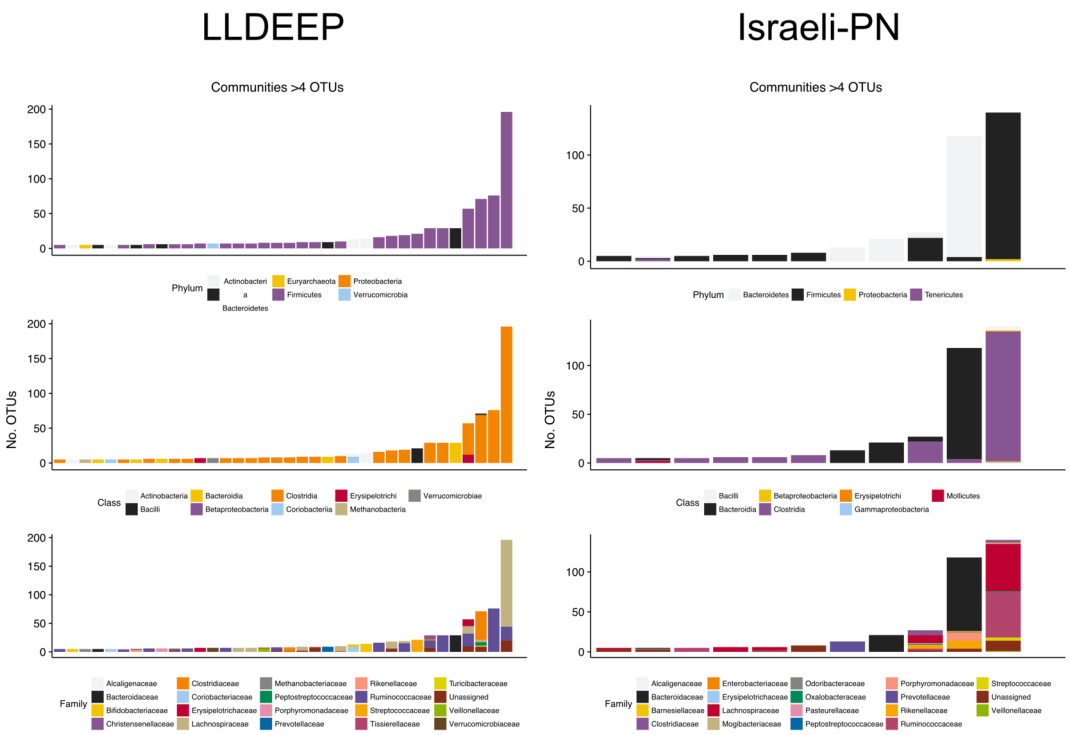


Figure E.2: Taxonomic summary plots for the communities in the LLDEEP and Israeli-PN data, as for TwinsUK in Figure 7.8.